When Do Annotator Demographics Matter? Measuring the Influence of Annotator Demographics with the POPQUORN Dataset

Jiaxin Pei
School of Information
University of Michigan
`pedropei@umich.edu`

David Jurgens
School of Information
University of Michigan
`jurgens@umich.edu`

Abstract

Annotators are not fungible. Their demographics, life experiences, and backgrounds all contribute to how they label data. However, NLP has only recently considered how annotator identity might influence their decisions. Here, we present POPQUORN (the **Po**tato-**P**rolific dataset for **Qu**estion-Answering, **O**ffensiveness, text **R**ewriting and politeness rating with demographic **N**uance). POPQUORN contains 45,000 annotations from 1,484 annotators, drawn from a representative sample regarding sex, age, and race as the US population. Through a series of analyses, we show that annotators' background plays a significant role in their judgments. Further, our work shows that backgrounds not previously considered in NLP (e.g., education), are meaningful and should be considered. Our study suggests that understanding the background of annotators and collecting labels from a demographically balanced pool of crowd workers is important to reduce the bias of datasets. The dataset, annotator background, and annotation interface are available at https://github.com/Jiaxin-Pei/potato-prolific-dataset.

1 Introduction

Supervised machine learning relies heavily on datasets with high-quality annotations and data labeling has long been an integral part of the machine learning pipeline (Roh et al., 2019). While recent large language models show promising performances on many zero-shot and few-shot NLP tasks (Bang et al., 2023), reinforcement learning with human feedback (RLHF), the core technology behind these models also heavily relies on large-scale and high-quality human annotations (Ouyang et al., 2022; Stiennon et al., 2020). Therefore, how to curate high-quality labeled datasets is one of the most important questions for both academia and industry.

Crowdsourcing is actively used as one of the major approaches to collect human labels for various NLP and ML tasks. Early studies on crowdsourcing NLP datasets suggest that crowd workers are able to generate high-quality labels even for relatively difficult tasks and with relatively low costs (Snow et al., 2008). However, other studies also suggest that collecting high-quality annotations from crowdsourcing platforms is challenging and requires rounds of iterations to create a reliable annotation pipeline (Zha et al., 2023).

Annotation quality has typically been measured by proxy through inter-annotator agreement (IAA) metrics like Krippendorff's α (Krippendorff, 2011) or Cohen's κ (Kvålseth, 1989). To attain higher IAA, researchers usually conduct pilot studies or rounds of annotator training to attain higher agreement among annotators. While such a method generally works in settings like part of speech tagging, the use of IAA as a proxy for quality implicitly assumes that the task has real ground truth and disagreements are mistakes. However, annotations for subjective tasks presents a far more challenging setting (Sandri et al., 2023); and as NLP and ML models are more frequently used in social settings where single true answer may not naturally exists, using IAA as the single metric for data quality can be problematic or can even create social harm. For example, Sap et al. (2021) studies how annotators' identity and prior belief affect their ratings on language toxicity and found significant gender and race differences in rating toxic language. Other studies also suggest that disagreement in annotations can also be due to the inherent contextual ambiguity (Jurgens, 2013; Poesio et al., 2019; Pavlick and Kwiatkowski, 2019) which can also be leveraged to improve the model performances (Uma et al., 2021).

Despite multiple studies on annotator background and disagreement (Wan et al., 2023; Orlikowski et al., 2023), a systematic study on

Task	Description	Data	Total Annotations	Number of Annotators	Instances	Average Labels per Instance
Offensiveness rating	Rate comment offensiveness using a 1-5 scale	Ruddit	13,036	262	1,500	8.7
Question Answering	Read a passage and answer a question through highlighting the text	SQuAD	4,576	459	1,000	4.6
Text rewriting / Style transfer	Read an email and revise it to make it sound more polite	Enron	2,346	257	1,429	1.6
Politeness Rating	Rate the politeness of an email using a 1-5 scale	Enron	25,042	506	3,718	6.7
POPQUORN			45,000	1,484	7,647	–

Table 1: POPQUORN contains 45,000 annotations from 1,484 participants from a representative sample regarding sex, age and race. Each annotator is paid $12 per hour as suggested by Prolific. POPQUORN covers four representative NLP tasks.

how annotator background influences different types of labeling tasks is still missing in the current literature. To address this gap, in this study, we present POPQUORN (the **Po**tato-**P**rolific dataset for **Qu**estion-Answering, **O**ffensiveness, text **R**ewriting and politeness rating with demographic **N**uance) a large dataset labeled by a US-population representative sample of annotators. POPQUORN contains 45,000 annotations for four diverse NLP tasks: offensiveness detection (classification/regression), questions answering (span identification), politeness style transfer (language generation) and politeness rating (classification/regression). All four tasks are annotated with a total of 1,484 annotators sampled from a representative sample regarding sex, age and race as the US population.

Through our analysis, we find that demographic background is significantly associated with people's ratings and performance on all four tasks— even for a more objective task such as reading comprehension. For example, people with higher levels of education perform better on the question-answering task and Black or African American participants tend to rate the same email as more polite and the same comment as more offensive than other racial groups. Our study suggests that demographic-aware annotation is important for various types of NLP tasks.

Overall our study makes the following four contributions. First, we create and release POPQUORN, a large-scale NLP dataset for four NLP tasks annotated by a representative sample of the US population with respect to sex, age, and race. Second, we analyze the annotations by different groups of annotators and found that various demographic backgrounds is significantly associated with people's rating of offensiveness, politeness as well as their performance on reading

comprehension. Third, in comparison with existing annotations from curated workers, we demonstrate that a general sample of Prolific participants can produce high-quality results with minimal filtering, suggesting the platform is a reliable source of quality annotations. All the annotations, annotator background information, and labeling interface are available at https://github.com/Jiaxin-Pei/potato-prolific-dataset.

2 Motivation

Individual and group differences are two of the most fundamental components of social sciences (Biggs, 1978). Social and behavioral sciences exist, in part, because of systematic human variations: if everyone were to behave in the same way, there would be no need to build theories and models to understand people's behaviors in different settings. As a special form of human task, data labeling is also subject to such a basic rule: different people may have different perceptions of various information and different performances on various language tasks. In this sense, while NLP researchers try to achieve a higher IAA, disagreement is a natural and integral part of any human annotation task (Leonardelli et al., 2021). Existing studies in this direction generally focus on building models that can learn from human disagreement (Uma et al., 2021) and some recent studies start to look at how annotators' identity and prior belief could affect their ratings in offensive language and hate speech (Sap et al., 2019, 2021). However, most of the existing studies only focus on selected dimensions of identities (e.g., gender) and on certain tasks (e.g., toxic language detection).

Our study aims at providing a systematic examination of how annotators' background affect their perception of and performances on various

language tasks. On the annotator side, we use a representative sample that matches the sex, age and race distribution of the US population. On the task side, we try to select tasks that are representative of common NLP tasks and with different degrees of difficulty, creativity, and subjectivity. Following this criterion, we selected four NLP tasks: (1) offensiveness detection, which is a relatively subjective task for classification and regression, (2) question answering, which is an objective task for span identification that is argued to test reading comprehension, (3) email rewriting, which requires creativity for a text generation task, and (4) politeness rating, which is also a subjective task for classification and regression.

3 Task 1: Offensiveness detection

Abusive or offensive language has been one of the most prominent issues on social media (Saha et al., 2023) and many existing studies tried to build datasets and models to detect offensive language (Malmasi and Zampieri, 2017; Yin and Zubiaga, 2021). Despite all the efforts on offensiveness detection, these models and datasets may have their own biases and during the creation of these datasets, annotators may introduce their own biases into their labels (Sap et al., 2019)—possibly marginalizing populations whose views differ from the majority. Indeed, Breitfeller et al. (2019) show that it was necessary to model the disparity between ratings from men and women annotators to identify gender-based microaggressions. However, most of the existing studies do not report the background of the annotators (Vidgen and Derczynski, 2020). Vidgen and Derczynski (2020) reviewed 63 offensiveness datasets and found that only 12 of them report detailed information about annotators.

To understand how annotator backgrounds (e.g. age, sex and race) affect their ratings on offensiveness, we re-annotated 1500 comments sampled from the Ruddit dataset (Hada et al., 2021) using 262 annotators from a representative sample from prolific.co.[1] In this section, we introduce the data sampling process, annotation task design, annotation result and then discuss how annotators' background affect their ratings of offensiveness.

3.1 Data and sampling

We use the Ruddit dataset (Hada et al., 2021) which contains 6,000 Reddit comments annotated using best-worst scaling (BWS; Flynn and Marley, 2014). Each comment is associated with an offensiveness score ranging from -1 to 1, computed from the BWS ratings. To select the subset we annotate, we remove comments that are shorter than 4 words or longer than 100 words, comments containing URLs as well as quote comments. Such a process led to 5,658 cleaned comments from the Ruddit dataset. We speculated that annotator background might be most influential in borderline cases, i.e., those not extremely offensive or inoffensive. Therefore, we use bucketed sampling based on the offensiveness score and we sample 10% from (-1, -0.5), 30% from (-0.5, 0), 50% from (0, 0.5) and 20% from (0.5, 1). Figure 1 shows the distribution of offensiveness scores before and after the sampling process. Our sampling process produced a subsample of comments with potentially more balanced offensiveness scores.

3.2 Task design

Each participant is presented with 50 comments and is asked to rate "Consider you read the above comment on Reddit, how offensive do you think it is?" using a 1-5 Likert scale where 1 means "Not offensive at all" and 5 means "Very offensive". Prior to annotating, each participant is shown an explicit warning about potentially seeing offensive content and has to answer a consent question before any comment is presented. When Prolific provides a demographically-representative sample, some information on the participants is provided. However, to ensure participants consent to have this information shared and reported as they themselves identify, we include a demographic and background screening question after the study is finished. Participants are shown an explanation for why demographic information was being asked for and were allowed to select "prefer not to disclose" if they wished.

To validate the annotation procedure, we conducted a pilot study with 8 participants. We used MACE (Hovy et al., 2013) to calculate the annotator competence score and ultimately removed one annotator with a competence score lower than 0.1. The annotators attain moderate IAA (Krippendorff's α=0.35), which is on par with existing studies on offensiveness labeling (Kang and Hovy,

---

[1] Prolific provides a service to request a sample of annotators with the same distribution of sex, age, and race as the US population using participants self-reported identities. We note that these demographic categories are based on the US Census questions in order to estimate a balanced sample.
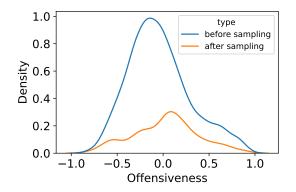
Figure 1: The offensiveness score in the Ruddit data before and in our subset after sampling. Positive scores denote offensive text.
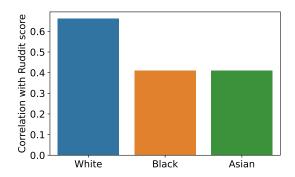

Figure 2: Correlation with the original Ruddit offensiveness score by race. Annotations by White participants have the highest correlation with the Ruddit score, while annotations by Asian and Black participants are significantly less correlated.

2021). We use POTATO (Pei et al., 2022) to set up the annotation website because of its integration with Prolific; Appendix Figure 8 shows the annotation interface.

3.3 Annotation result

The full annotation process collected 13,036 annotations from 262 participants and each comment received 8.7 annotations on average. The medium time of finishing 50 annotations is 13 minutes. Krippendorff's α=0.29, showing moderate to low agreement among annotators. However, the overall correlation between the averaged annotations and the original Ruddit score is 0.67, suggesting that, on average, the judgments largely matched those of the original dataset. Participants were highly open to sharing their demographics, with over 95% filling out the questionnaire.

	Coef.	Std.Err.	z	P> |z|
intercept	1.998	0.048	41.259	0.000
gender: Non-binary	-0.235	0.060	-3.890	0.000
gender: Woman	-0.022	0.020	-1.065	0.287
race: Black or African American	0.184	0.045	4.124	0.000
race: Hispanic or Latino	-0.405	0.078	-5.174	0.000
race: White	-0.104	0.038	-2.758	0.006
age: 25-29	-0.185	0.043	-4.268	0.000
age: 30-34	-0.165	0.041	-4.071	0.000
age: 35-39	-0.142	0.040	-3.525	0.000
age: 40-44	-0.037	0.043	-0.860	0.390
age: 45-49	-0.087	0.044	-1.979	0.048
age: 50-54	-0.141	0.046	-3.077	0.002
age: 54-59	0.001	0.039	0.025	0.980
age: 60-64	0.309	0.050	6.163	0.000
age: >65	0.117	0.042	2.755	0.006
education: College degree	-0.015	0.023	-0.660	0.509
education: Graduate degree	0.052	0.029	1.801	0.072

Table 2: Mixed-effect regression results showing the influence of annotator demographics on their offensiveness rating, controlling for the item being rated. Reference categories are Gender: Men, Race: Asian, Age: 18-25, and Education: High school degree.

3.4 Does annotator background affect offensiveness rating?

To understand the influence of annotator background on offensiveness ratings, we ran a linear mixed-effect model to predict the offensiveness rating with gender, age, race, and educational background, controlling each instance as the random effect. By controlling for each instance, we control for differences in the relative levels of offensiveness between instances, which allows us to study deviations from a mean judgment. Categories that are too rare in the data are removed from the regression (e.g. only 1 annotator chooses the "other" category for education). 16 annotators are dropped from this process.

Gender Do men and women have different ratings for offensiveness? Surprisingly, while some existing studies suggest that men and women may have different ratings of toxic language (Sap et al., 2021), we found no statistically significant difference between men and women. However, participants with non-binary gender identities tend to rate messages as less offensive than those identifying as men and women.

Age People older than 60 tend to perceive higher offensiveness scores than middle-aged participants. It is possible that older people are more sensitive to offensive language and they are less exposed to the language style of Reddit comments. Younger individuals are known to avoid swearing in the presence of older individuals but not among peers

(Fägersten, 2012, p. 111) and that younger individuals tend to use stronger swearing (Gauthier and Guille, 2017), which supports the idea that inter-generation norms may lead to differences in the perception of toxicity.

Race   We found significant racial differences in offensiveness rating: Black participants tend to rate the same comments with significantly more offensiveness than all the other racial groups. In this sense, classifiers trained on data annotated by White people may systematically underestimate the offensiveness of a comment for Black and Asian people.

Education   No significant differences were found with respect to annotator education, though the relatively small effect for those with graduate degrees does approach significance.

3.5   Are Ruddit annotations closer to perceptions of people in certain ethnicity groups?

We calculate the aggregated score of each racial group and calculate the overall correlation with the Ruddit offensiveness score. As shown in Figure 2, scores by White annotators are highly correlated with the Ruddit annotations (Pearson's r=0.66), while the scores by Black, and Asian annotators are only moderately correlated with the Ruddit score (Pearson's $r \approx$0.4), suggesting that the Ruddit annotations are more likely to have been done by White annotators.

4   Task 2: Question Answering

Question Answering/Reading comprehension is one of the most fundamental tasks of NLP (Rogers et al., 2023) and SQuAD has been actively used by the research community to evaluate the performance of their models on question answering (QA) as a form of reading comprehension (Rajpurkar et al., 2016, 2018). To evaluate crowd workers' ability to complete QA tasks and study whether participants' background is associated with different performances, we build the second task as part of the POPQUORN.

4.1   Data and sampling

We use SQuAD 2.0 (Rajpurkar et al., 2018) as it also contains unanswerable questions and can pose external challenges to the annotators compared with SQuAD 1.0. In SQuAD 2.0, each passage can contain multiple questions. We sample 1000 unique passages and questions from the SQuAD 2.0 dataset. The final sampled dataset contains 695 questions with correct answers and 305 unanswerable questions.

4.2   Annotation task design

We recruit participants from a US-population representative sample (with respect to sex, age, and ethnicity) on Prolific. Each annotator is assigned with 10 passage and question pairs and is paid $12 per hour for their participation. At the end of the study, their demographic information is collected through an after-study survey. Besides the question-answering schema, we also ask participants to self-report the difficulty of their questions as task difficulty might be associated with disagreement (Uma et al., 2021). Appendix Figure 9 shows the annotation interface for this task.

4.3   Annotation result

4,576 annotations are collected from 459 annotators. Each question received 4.6 annotations on average (similar to the SQuAD data where on average 4.8 answers are collected for each question). We use a similar strategy as Rajpurkar et al. (2016) to aggregate the answers for each question: choose the majority answer and use the shorter version if there is a tie. We use the evaluation script provided by SQuAD to calculate the token-level precision, recall and F1 score for each answer. The aggregated answers achieve 0.75 F1, 0.72 precision, and 0.79 recall.

We manually examined a sample of human errors and we found that the crowd workers are mostly able to identify the correct answer but may use a larger span, which leads to higher recall but lower precision. More specifically, we annotated 50 instances where the F1 score is lower than 1 and found that for all these instances, at least one annotator is able to answer it correctly. Moreover, the SQuAD groundtruth is only correct in 12 out of 50 (24%) instances and for 8 instances (16%), the crowdworkers are able to identify the correct answer where the SQuAD groundtruth is incorrect. We found 2 out of 50 (4%) instances that both SQuAD and our crowdworkers didn't answer the question correctly.

What demographic factors influence answer accuracy?   To study the connection between demographic background and performance on the read-

	Coef.	Std.Err.	z	P> |z|
Intercept	0.580	0.032	18.238	0.000
gender: Non-binary	0.008	0.036	0.233	0.816
gender: Woman	-0.031	0.013	-2.392	0.017
race: Black or African American	-0.092	0.032	-2.847	0.004
race: Hispanic or Latino	-0.149	0.038	-3.874	0.000
race: White	-0.062	0.028	-2.242	0.025
age: 25-29	0.012	0.029	0.404	0.686
age: 30-34	0.040	0.027	1.491	0.136
age: 35-39	-0.050	0.028	-1.779	0.075
age: 40-44	0.072	0.028	2.567	0.010
age: 45-49	0.079	0.027	2.903	0.004
age: 50-54	0.116	0.029	3.938	0.000
age: 54-59	0.072	0.027	2.697	0.007
age: 60-64	0.002	0.027	0.060	0.952
age: >65	0.008	0.026	0.311	0.756
education: College degree	0.027	0.015	1.824	0.068
education: Graduate degree	0.060	0.018	3.382	0.001

Table 3: Mixed-effect regression results showing the influence of annotator demographics on their performance at question answering (as measured by F1 score), controlling for the item being rated. Reference categories are Gender: Men, Race: Asian, Age: 18-25, and Education: High school degree.

ing comprehension task, we run a mixed effect model as in §3.4 with variables for gender, age, education, and ethnicity as fixed effects and the instance as the random effect. Despite the task being largely objective, accuracy at question answering varied relative to annotator background, as shown in Table 3. The largest effects were seen with race and age variation, with a smaller effect for education. While the root causes of this performance disparity cannot be directly tested from our survey, two notable general trends are worth mentioning.

First, the performance differences mirror known disparities in education and economic opportunities for minorities compared with their White male peers in the US.[2] Multiple studies have shown how structural forces have led to lower levels of reading abilities by race (Dixon-Román et al., 2013; Merolla and Jackson, 2019) and socioeconomic status (Merz et al., 2020).

Second, the trend for age matches known results showing a moderate increase in reading ability with age (Pfost et al., 2014; Locher and Pfost, 2020). Further, Locher and Pfost (2020) also note that individuals in professions that require reading also have better reading comprehension, which we view as a potential contributor to the performance increase seen from annotators with graduate degrees who are more likely to have such professions.
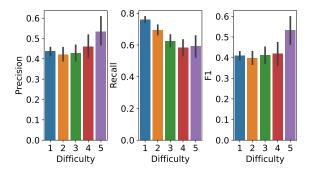
Figure 3: Questions rated with lower difficulty are generally associated with higher Recall. However, when people use the highest difficulty score, people generally perform better as measured by precision, suggesting that people tend to be more selective about their answers when they perceive that the task is difficult.

Is self-reported difficulty associated with participant performance? During the study, participants are also asked to rate "How difficult do you think this question is" on a 1-5 likert scale where 1 means not difficult at all and 5 means very difficult. Figure 3 shows the overall F1, recall and precision score and the difficulty rated by each participant. We found that when people report lower difficulty, their recalls tend to be higher, suggesting that they are better able to identify the potential span of the answer. However, perceived difficulty is also associated with increased precision. Multiple mechanisms might explain this pattern: it is possible that difficult questions require a more specific answer. It is also possible that people may be more cognitively focused to solve the challenge when they perceive the task is more difficult.

5 Task 3: Politeness rewriting

Politeness is one of the most prominent social factors in interpersonal communication (Brown et al., 1987). The NLP community has built computational models for predicting politeness scores and built models to generate polite text in different settings (Danescu-Niculescu-Mizil et al., 2013; Madaan et al., 2020; Porayska-Pomsta and Mellish, 2004). However, few resources exist with human-authored examples of pairs of original and style-transferred texts for politeness. Therefore, to test the crowdworker's ability to generate open-domain text for style-transfer tasks, we recruit participants from Prolific to rewrite emails from the Enron dataset as part of POPQUORN.
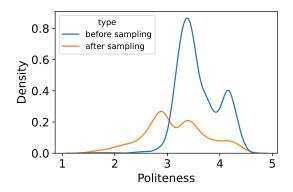
Figure 4: The inferred politeness score of emails in the original Enron dataset and in our subset after sampling. Higher scores indicate higher degrees of politeness.

5.1 Data and samples

We use the Enron email dataset (Shetty and Adibi, 2004) which contains approximately 500,000 emails from senior management executives at the Enron Corporation. We first extract the main body of the emails and then we remove emails that are too long (larger than 100 words), too short (shorter than 8 words), containing URLs, containing more than 10 numbers or were automatically generated by systems. This preprocessing lead to 84,066 remaining emails. We use `politenessr` [3] to infer the politeness score of each email. As most of the emails are relatively polite in the dataset, to draw a more balanced sample for annotation, we use bucket sampling and sample 50% from (1,3), 40% from (3,4), and 10% from (4,5). The final dataset used for annotation contains 1000 emails. Figure 4 shows the distribution of politeness score after bucket sampling. The sampled emails contain more emails with lower inferred politeness scores than the original Enron dataset.

5.2 Annotation task

In the annotation task, each annotator is presented with 10 emails and asked to "rewrite the email to make it sound more polite in a work setting". Appendix Figure 10 shows the annotation interface for this rewriting task. We conduct a pilot study with 18 participants to validate the annotation procedure. The pilot study attained 180 annotations for 150 emails and the average editing distance is 102, suggesting that the annotators are making substantial changes to the original message. Politeness of the

[3] The model is accessible at https://github.com/wujunjie1998/Politenessr and was trained on politeness data from Danescu-Niculescu-Mizil et al. (2013) and Wang and Jurgens (2018).
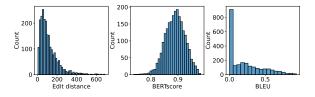

Figure 5: Measures comparing the original and revised emails show that the revisions are still very semantically similar (high BERTScore) but the form of the content has been substantially changed (high edit distance and low BLEU score).
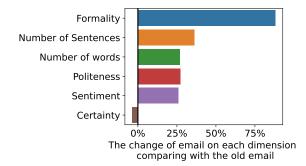

Figure 6: The revised emails have 36% more sentences and 26% more words than the original emails. Moreover, the revised emails are 88% more formal, 27% more polite, 25% more positive, and 3% less certain than the old emails, suggesting that the participants are making substantial changes to make the email more polite.

revised email increases by 0.53 on average when compared with the original emails, suggesting that the revised emails are much more polite.

5.3 Full Annotation Results

The final politeness rewriting dataset contains 2,346 emails written by 257 participants drawn from a US population representative sample (regarding sex, age, and race). In the final dataset, we remove the revised emails if they are shorter than 7 words or if the edit distance is lower than 5 (79 out of 2376 emails are removed). As shown in Figure 6, the overall politeness increase 27% compared with the original emails, suggesting that the rewritten emails are significantly more polite than the original ones. The revised emails are more positive[4], more formal[5] and less certain[6] comparing with the original emails. To achieve these changes, annotators substantially changed the emails, with

[4] https://huggingface.co/Seethal/sentiment_analysis_generic_dataset
[5] https://huggingface.co/s-nlp/roberta-base-formality-ranker
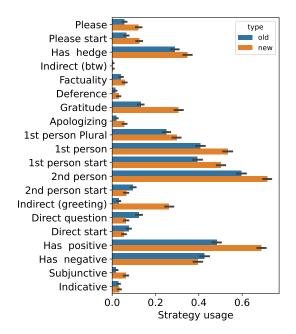[6] https://pypi.org/project/certainty-estimator/ (Pei and Jurgens, 2021)

Figure 7: Annotators adopt a wide range of politeness strategies.

an average editing distance of 112; this indicates that changes were mostly not perfunctory, small edits.

Despite these changes to the tone and style of the emails, annotators kept the meaning largely consistent. Figure 5 shows the distribution of edit distance, BERTscore (Zhang et al., 2019) and BLEU, with the latter two being proxies for the interpretation or meaning of the email content. The BERTscore for the emails is generally above 0.8, suggesting that the revised email are able to retain the meaning of the original content. On the other hand, most of the BLEU scores are lower than 0.2, suggesting that the participants are able to make changes to the original content while keeping its meaning.

Did annotators use a diversity of strategies for increasing politeness—or did they just add "please" to every sentence? To further understand changes annotators made to the original emails, we analyzed the politeness strategies using ConvoKit (Chang et al., 2020) and compared the strategies' prevalence in both revised and original emails. As shown in Figure 7, annotators adopt a wide range of politeness strategies (Danescu-Niculescu-Mizil et al., 2013). The usage of "please" in a sentence does increase (as expected), and we see a larger increase in strategies such as expressing gratitude, use of positive words, and indirect greetings. Together, this variation suggests that the revisions capture more

natural variation in writing and are not artificial revisions driven by task design or speed incentives.

6 Task 4: Politeness Rating

To validate the email rewriting results from Task 3, we perform a follow-up participant recruitment to rate the politeness of the original and revised emails. A recent study shows that the nativeness of the annotators affect the IAA of politeness rating (Srinivasan and Choi, 2022). In this study, we release the full backgrounds of annotators and hope this dataset helps promote future studies on politeness prediction and to understand how people with different backgrounds perceive politeness.

6.1 Annotation setup

We use 1,372 emails from the original Enron dataset and 2,346 emails rewritten by the participants. Annotators are asked to rate "Consider you read this email from a colleague, how polite do you think it is?" using a 1-5 Likert scale where 1 means "not polite at all" and 5 means "Very polite". Each annotator is presented with 50 emails in a random order and on average each email is annotated by 6.7 annotators. Appendix Figure 11 shows the interface of this annotation task. We ran one pilot study with 8 annotators and each annotator is presented with 50 emails. The overall Krippendorff's α is 0.43, suggesting moderate inter-annotator agreement and is reasonable for such a subjective task.

6.2 Full annotations

Our final politeness rating dataset contains 25,042 annotations from 506 annotators. Each email receives 6.7 annotations on average. The overall Krippendorff's α is 0.43, indicating moderate to low inter-annotator agreement. The overall politeness rating is 2.8 and 3.6 for original and revised emails, suggesting that the revised emails are perceived as more polite than the original emails, which correlates with the previous result.

6.3 Does annotator background affect politeness rating?

We ran a linear mixed-effect model to predict the politeness rating with gender, age, race, and education, controlling each instance as the random effect, similar to previous setups. Table 4 shows the regression results.

Gender We found that women rate messages as less polite, though the effect size is relatively small

	Coef.	Std.Err.	z	P> |z|
Intercept	3.167	0.035	89.497	0.000
gender: Non-binary	-0.048	0.042	-1.149	0.250
gender: Woman	-0.042	0.014	-3.116	0.002
race: Black or African American	0.192	0.032	6.105	0.000
race: Hispanic or Latino	0.057	0.036	1.607	0.108
race: White	0.060	0.027	2.212	0.027
age: 25-29	0.291	0.030	9.630	0.000
age: 30-34	0.078	0.028	2.764	0.006
age: 35-39	0.169	0.031	5.376	0.000
age: 40-44	0.137	0.029	4.704	0.000
age: 45-49	0.296	0.031	9.677	0.000
age: 50-54	0.305	0.030	10.275	0.000
age: 54-59	0.198	0.029	6.717	0.000
age: 60-64	0.249	0.029	8.623	0.000
age: >65	0.209	0.028	7.508	0.000
education: College degree	-0.145	0.015	-9.394	0.000
education: Graduate degree	-0.135	0.020	-6.837	0.000

Table 4: Mixed-effect regression results showing the influence of annotator demographics on their politeness ratings, controlling for the item being rated. Reference categories are Gender: Men, Race: Asian, Age: 18-25, and Education: High school degree.

compared with other demographic dimensions.

Age   Compared with the youngest segment in our sample (Ages 18-25), all older segments were more likely to give a higher politeness rating.

Race   We found significant racial differences in politeness rating. Relative to Asian peers, Black participants rated messages as more polite, with a small positive effect for White peers. No significant result was seen for annotators identifying as Hispanic or Latino. Given known differences in the cultural perceptions of politeness (Troutman, 2010; Brown, 2015; Rodríguez-Arauz et al., 2019), these differences suggest systematic variation in the rating that would otherwise be treated as disagreement, rather than valid, culturally-situated judgments.

Educational Background   As shown in Table 4, participants with more education (a graduate or college degree) tend to rate the same email with less politeness than those with a high school degree. Education is strongly correlated with socioeconomic status and with that status typically comes increased social standing. While multiple works have shown how individuals modify their speech with respect to power/status differences between speaker and recipient (e.g., Brown et al., 1987; Wang, 2021), we believe our result offers a valuable new insight to how individuals with different status view the *same* message. Our results suggest that higher-status (more educated) individuals are

7   Discussion

High-quality annotated data has been one of the primary driving factors of NLP and ML. While some studies try to look at improving data quality through analyzing disagreements among annotators, systematic studies of how annotators' background affects crowdsourcing results remain rare. In this paper, we create a new NLP dataset labeled by annotators from a US-representative sample regarding sex, age and race. We re-annotated the Ruddit offensiveness dataset and found that the offensiveness is strongly correlated with annotations by White participants, while the correlation between the Ruddit offensiveness score and annotations by participants from other racial groups are only 0.41, suggesting that the Ruddit dataset might largely reflect the views of White annotators of what content is offensive. As people from other cultures may perceive the same comment with a lower or higher degree of offensiveness, classifiers trained on a dataset annotated by White participants could pose risks for many people. Such an issue becomes increasingly important as both the industry and research community are trying to align the values of LLMs with human beings through instruction tuning.

8   Conclusion

Who annotates your data matters. Across four annotation tasks, we show that an annotator's background influences their decisions, across multiple annotation tasks with different degrees of subjectivity. In more subjective tasks, these differences in decisions are not mistakes but rather valid differences in views. Our results underscore that NLP papers that curate datasets must consider whose voices appear in their datasets, as these ultimately decide whose voice are captured in models trained on the data. Indeed, by comparing our annotations with those from the existing annotated datasets, we show that the existing annotated dataset might be annotated by a demographically-biased group of annotators. To support work in this modeling demographic-aware and socially-responsible NLP, we release POPQUORN with 45,000 annotations on four NLP tasks by nearly 1.5K annotators.

Acknowledgments

Ethical Implications

Collecting background information about annotators can be sensitive and have ethical implications. In our study, we follow best practices when asking about demographic information (Spiel et al., 2019) and always allow participants to choose "Prefer to not disclose" and provide external options for them to self-describe identities. Understanding how different groups of people perceive social information in language and perform different tasks is important when NLP models are applied in more and more social applications. We believe that through carefully designed procedures to collect background information of annotators along with the data annotation, we will be able to build better NLP and ML models that could better serve different groups of people and reduce potential social harm.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

John B Biggs. 1978. Individual and group differences in study processes. *British Journal of Educational Psychology*, 48(3):266–279.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674.

Penelope Brown. 2015. Politeness and language. In *The International Encyclopedia of the Social and Behavioural Sciences (IESBS),(2nd ed.)*, pages 326–330. Elsevier.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259.

Ezekiel J Dixon-Román, Howard T Everson, and John J McArdle. 2013. Race, poverty and sat scores: Modeling the influences of family income on black and white high school students' sat performance. *Teachers College Record*, 115(4):1–33.

Kristy Beers Fägersten. 2012. *Who's swearing now? The social aspects of conversational swearing*. Cambridge Scholars Publishing.

Terry N Flynn and Anthony AJ Marley. 2014. Best-worst scaling: theory and methods. In *Handbook of choice modelling*, pages 178–201. Edward Elgar Publishing.

Michael Gauthier and Adrien Guille. 2017. Gender and age differences in swearing. *Advances in swearing research: New languages and new contexts*, pages 137–156.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for english reddit comments. *arXiv preprint arXiv:2106.05664*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562.

Dongyeop Kang and Eduard Hovy. 2021. Style is not a single variable: Case studies for cross-stylistic language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2376–2387.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Tarald O Kvålseth. 1989. Note on cohen's kappa. *Psychological reports*, 65(1):223–226.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *arXiv preprint arXiv:2109.13563*.

Franziska Locher and Maximilian Pfost. 2020. The relation between time spent reading and reading comprehension throughout the life course. *Journal of Research in Reading*, 43(1):57–77.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. *arXiv preprint arXiv:2004.14257*.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

David M Merolla and Omari Jackson. 2019. Structural racism as the fundamental cause of the academic achievement gap. *Sociology Compass*, 13(6):e12696.

Emily C Merz, Elaine A Maskus, Samantha A Melvin, Xiaofu He, and Kimberly G Noble. 2020. Socioeconomic disparities in language input are associated with children's language-related brain structure and reading skills. *Child development*, 91(3):846–860.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics. *arXiv preprint arXiv:2306.11559*.

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.

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

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Jackson Sargent, Apostolos Dedeloudis, and David Jurgens. 2022. Potato: The portable text annotation tool. *arXiv preprint arXiv:2212.08620*.

Jiaxin Pei and David Jurgens. 2021. Measuring sentence-level and aspect-level (un) certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9959–10011.

Maximilian Pfost, John Hattie, Tobias Dörfler, and Cordula Artelt. 2014. Individual differences in reading development: A review of 25 years of empirical research on matthew effects in reading. *Review of educational research*, 84(2):203–244.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789.

Kaśka Porayska-Pomsta and Chris Mellish. 2004. Modelling politeness in natural language generation. In *Natural Language Generation: Third International Conference, INLG 2004, Brockenhurst, UK, July 14-16, 2004. Proceedings*, pages 141–150. Springer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Gloriana Rodríguez-Arauz, Nairán Ramírez-Esparza, Adrián García-Sierra, Elif G Ikizer, and María José Fernández-Gómez. 2019. You go before me, please: Behavioral politeness and interdependent self as markers of simpatía in latinas. *Cultural Diversity and Ethnic Minority Psychology*, 25(3):379.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347.

Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. 2023. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11):e2212270120.

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.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Jitesh Shetty and Jafar Adibi. 2004. The enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4(1):120–128.

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.

Katta Spiel, Oliver L Haimson, and Danielle Lottridge. 2019. How to do better with gender on surveys: a guide for hci researchers. *Interactions*, 26(4):62–65.

Anirudh Srinivasan and Eunsol Choi. 2022. Tydip: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738.

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

Denise Troutman. 2010. Attitude and its situatedness in linguistic politeness. *Poznań Studies in Contemporary Linguistics*, 46(1):85–109.

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.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *arXiv preprint arXiv:2301.05036*.

Yi Wang. 2021. The price of being polite: politeness, social status, and their joint impacts on community q&a efficiency. *Journal of Computational Social Science*, 4:101–122.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

I can only try to change your view by offering problems that may be larger - I just can't downplay the importance of your point. 1. The method we use to vote (referred to as "first past the post" I believe) will always lead to a two party system. If we change the method used for voting, we would have more choices; it would no longer be us vs. them. 2. It is human nature to stay away from perspectives we disagree with. Viewing the other side as the enemy is just a small part of this problem.

Consider you read the above comment on Reddit, how offensive do you think it is?

Not offensive at all ○ ○ ○ ○ ○ Very offensive

[Move backward] [Move forward]

Figure 8: Annotation interface for the offensiveness rating task.

Medical facilities in Mali are very limited, and medicines are in short supply. Malaria and other arthropod-borne diseases are prevalent in Mali, as are a number of infectious diseases such as cholera and tuberculosis. Mali's population also suffers from a high rate of child malnutrition and a low rate of immunization. An estimated 1.9 percent of the adult and children population was afflicted with HIV/AIDS that year, among the lowest rates in Sub-Saharan Africa. An estimated 85–91 percent of Mali's girls and women have had female genital mutilation (2006 and 2001 data).

Question: Malians suffer from malnutrition and low rates of what type of medical need?

Answer the question through highlighting the paragraph

☐ Answer

☐ No answer is given in this document

How difficult do you think this question is?

Not difficult at all ○ ○ ○ ○ ○ Very difficult

[Move backward] [Move forward]

Figure 9: Annotation interface for the SQuAD reading comprehension task

yes. you need to come over more.

Your collegue is trying to draft an email and is asking for your help.
Please help to rewrite the above message to make it more polite

Move backward   Move forward

Figure 10: Annotation interface for the email rewriting task

Lauren,
Please help with who the appropriate person
is to complete and submit w-8
forms for ENRON INDIA GDR HOLDINGS LTD.
Thanks. Sheila

Consider you read this email from a colleague, how polite do you think it is?

Not polite at all ○ ○ ○ ○ ○ Very polite

Move backward   Move forward

Figure 11: Annotation interface for the politness rating task.