On the Role of Speech Data in Reducing Toxicity Detection Bias

Samuel J. Bell[*†1], Mariano Coria Meglioli[†1], Megan Richards[†2], Eduardo Sánchez[†1,3], Christophe Ropers[1], Skyler Wang[4], Adina Williams[1], Levent Sagun[†‡1], Marta R. Costa-jussà[†‡1]

[1]Meta FAIR [2]New York University

[3]University College London [4]McGill University

[*]First author [†]Core contributor [‡]Senior author

Correspondence: sjbell@meta.com

Abstract

Text toxicity detection systems exhibit significant biases, producing disproportionate rates of false positives on samples mentioning demographic groups. But what about toxicity detection in *speech*? To investigate the extent to which text-based biases are mitigated by speech-based systems, we produce a set of high-quality group annotations for the multilingual MuTOX dataset, and then leverage these annotations to systematically compare speech- and text-based toxicity classifiers. Our findings indicate that access to speech data during inference supports reduced bias against group mentions, particularly for ambiguous and disagreement-inducing samples. Our results also suggest that improving classifiers, rather than transcription pipelines, is more helpful for reducing group bias. We publicly release our annotations and provide recommendations for future toxicity dataset construction.

Content warning: This paper contains toxic language that readers may find offensive or upsetting.

1 Introduction

With the growing prevalence of machine learning systems capable of processing and generating speech, there is rising interest in speech-aware toxicity detection (Costa-jussà et al., 2024; Ghosh et al., 2022; Nandwana et al., 2024; Liu et al., 2024). Traditional cascaded approaches to speech toxicity detection use automated speech recognition (ASR) to convert speech to text, before applying a standard text classifier. This strategy has two main issues. First, it eliminates rich prosodic and contextual information present in speech, which could degrade model performance. Second, text-based toxicity detection systems are well known to exhibit significant biases against minoritized groups (Dixon et al., 2018; Borkan et al., 2019). For instance, many systems are more likely to consider African American English (AAE) as toxic (Resende et al., 2024), while others denote the mere mention of identities such as "gay" and "lesbian" as toxic (Dias Oliva et al., 2021). Often, these issues are attributed to biases in the training data. Because minoritized communities are overwhelmingly the subject of online toxicity (Dixon et al., 2018; Borkan et al., 2019), classifiers misinterpret *benign* group mentions as toxic, producing a disproportionate rate of false positives for marginalized groups (Dixon et al., 2018). Given these limitations, recent research has sought to develop toxicity classifiers that operate directly on speech.

In this work, we perform a systematic comparison of speech-based and cascaded text-based toxicity detection systems. Specifically, we hypothesize that access to speech audio provides useful contextual information, which could reduce false positives. To investigate this, we produce a new set of annotations for a multilingual speech toxicity dataset, MuTOX (Costa-jussà et al., 2024), annotating for both toxicity and group mentions while also correcting automated transcripts. To ensure consistent and accurate data, annotations were performed by the authors using a rigorous multi-stage process of cross-checking and discussion.

We leverage these annotations to produce critical new insight into both the efficacy and biases of speech-based and text-based toxicity detection models. **Our work reveals that incorporating speech data at inference time improves performance and reduces false positives on samples mentioning group identities, and eliminates false positives on ambiguous samples.** Furthermore, we find that this bias is not the result of transcription error, but of the classifier itself. We make our annotations publicly available to facilitate future research into the fairness and efficacy of speech-based toxicity detection.[1]

[1]https://fb.me/mutox-group-annotations

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1454–1468

Contributions. To summarize our main contributions, we:

1. Generate and release 1954 group annotations for speech toxicity detection fairness evaluations in English and Spanish;
2. Compare text- and speech-based toxicity detection systems, including detailed investigation of performance on ambiguous samples;
3. Isolate the role of transcription failure in text-based toxicity classifiers;
4. Provide extensive analysis of the challenging ambiguity of toxicity annotation in speech.

2 Background and related work

2.1 Bias in toxicity detection

Toxicity detection systems have long been known to exhibit significant biases (see Garg et al. 2023 for a review). One major issue is the over-representation of certain identity markers in toxicity detection training data, often correlated with toxic content (Dixon et al., 2018). For instance, models tend to conflate group mentions with toxicity, particularly for groups frequently targeted online, such as women, LGBTQ+ individuals, and minoritized racial, ethnic, or religious groups (Park et al., 2018; Borkan et al., 2019; Dias Oliva et al., 2021). Models explicitly designed to detect anti-group bias also incorrectly associate group mentions with toxicity (Sahoo et al., 2022), unable to distinguish the *use* of a term from a *mention* (Gligoric et al., 2024). Understanding how group mentions also bias speech toxicity classifiers is the key motivation of this work.

Toxicity classifiers have also been found to exhibit significant bias against AAE (Resende et al., 2024), partly due to annotator biases (Sap et al., 2022; Goyal et al., 2022). Racial bias has similarly been observed in hate speech detection, which also suffers from the challenge of disambiguating genuinely hateful from reappropriated words (Davidson et al., 2019; Sap et al., 2019).

Our work draws inspiration from Civil Comments (Borkan et al., 2019), a text toxicity dataset with group annotations. However, to better handle ambiguous cases, we opted to produce annotations ourselves rather than rely on crowd workers.

2.2 Speech toxicity detection

There is increasing interest in toxicity detection for speech data (Nandwana et al., 2024; Liu et al., 2024). The straightforward approach for constructing a speech-based toxicity detection system is a multi-stage pipeline, comprising an ASR stage followed by a text toxicity classification stage (Barrault et al., 2025). Alternatively, models that operate directly on speech (e.g. Costa-jussà et al. 2024) typically utilize self-supervised speech encoders trained on large volumes of speech data, including wav2vec (Baevski et al., 2020), WavLM (Chen et al., 2022), and SONAR (Duquenne et al., 2023). Prior work in speech profanity detection suggests that models benefit from access to "audio properties like pitch, emotions, [and] intensity" (Gupta et al., 2022, p. 4).

While there are both monolingual (Ghosh et al., 2022) and multilingual (Gupta et al., 2022; Costa-jussà et al., 2024) speech toxicity datasets, none are annotated with group information, precluding detailed analysis of bias against group mentions.

2.3 Bias in speech systems

Speech systems more broadly have been shown to exhibit biases in a range of contexts. For example, speech-based machine translation systems exhibit gender bias, such as by making gendered assumptions when translating between languages with and without grammatical gender (Costa-jussà et al., 2022). The same phenomenon is present in speech-enabled large language models (LLMs), though its severity appears to be language-specific (Lin et al., 2024).

Our work is closely connected to research exploring the biases of both ASR and self-supervised speech encoders such as SONAR (upon which MuTOX is based). Due to factors such as data imbalance (Garnerin et al., 2019), ASR systems can exhibit gender bias (Tatman, 2017) and accent bias (Feng et al., 2021), ultimately producing lower quality transcripts for certain groups of speakers. SSL speech encoders also exhibit biases with respect to accent, age, and nationality (Lin et al., 2024), though in contrast to ASR systems the composition of the pretraining data appears to have a limited effect (Boito et al., 2022; Meng et al., 2022). While speech *data* may provide useful context that could reduce bias, speech *pipelines* may add biases of their own, motivating our comparative study of text- and speech-based approaches.

Table 1: Selected samples with corrected transcripts and new toxicity and group annotations. English translations in gray. See appendix E for corresponding MuTOX IDs.

	Transcript	Toxic	Group
EN-1	"The Palestinian people does not exist"	Yes (hate speech)	Racial or ethnic groups
EN-2	"I'm gonna have sex with this guy"	No	Gender identities
ES-1	"Yo creo que la raza humana en general es una raza de mierda" "I believe that the human race in general is a shitty race"	Yes (profanity)	-
ES-2	"Él era una persona muy mala, mató a muchos judíos" "He was a very bad person, killed many Jews"	No	Religious groups

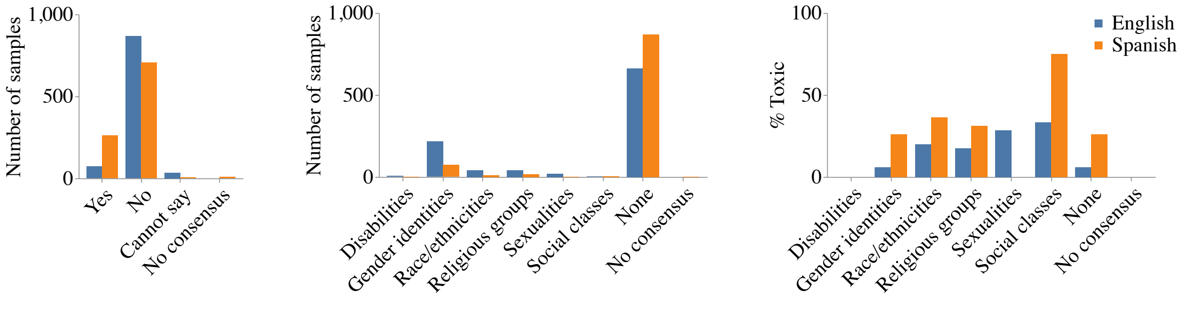


Figure 1: (a) Number of samples marked as toxic ("Yes"), not toxic ("No"), impossible to decide ("Cannot say"), or where annotators could not reach consensus ("No consensus") for English and Spanish. (b) Number of samples marked as mentioning or referring to a group. (c) Percentage of samples per group marked as toxic.

3 Annotating MuTOX

The foundational contribution of this work is a new, high-quality set of annotations for the MuTOX test partition, allowing us to evaluate classifier bias against group mentions. We believe this represents the first fairness audit dataset for multilingual speech toxicity detection.

3.1 The MuTOX dataset

MuTOX (Costa-jussà et al., 2024) is a large-scale, multilingual speech toxicity dataset covering 30 languages. Each audio sample is accompanied by a text transcript produced by an open-source ASR model (Radford et al., 2023). For annotation tractability, we focus only on the English and Spanish test partitions, covering a total of 1954 samples.

3.2 Stage 1: Initial annotation

We asked three annotators per language (all core contributors to this paper; see appendix A) to annotate the MuTOX test set. The annotators were all native-level proficient and spanned multiple language varieties (such as British and American English) to capture variety-specific interpretations. Annotators used LabelStudio (Tkachenko et al., 2020) with a custom interface (see appendix B) to annotate for toxicity, group mentions, and automated transcript correctness.

Toxicity. For toxicity, annotators were asked "Does the audio contain toxicity?" and presented with options for 'Yes," "No," or "Cannot say," the latter indicating that the audio was unclear, truncated, or context-dependent. Annotators were instructed to use the toxicity definition from the original MuTOX annotation guidelines (see appendix C), which defines toxicity as language which is "typically considered offensive, threatening or harmful." This includes profanities and language related to physical violence, bullying, pornography, or hate speech.

Group mentions. For group annotation, annotators were asked "Does the audio mention, or refer to (either explicitly or implicitly), any of the following?" to which they could respond with one or more of "Gender identities," "Sexualities," "Religious groups," "Racial or ethnic groups," "Disabilities," "Social classes or socio-economic statuses," or "None of the above." If any group was selected, annotators were asked then a follow-up about which specific group was mentioned. For example, in the case of gender identities, they were asked "Which gender identities are mentioned or referred to?" with predefined options: "Female, woman or girl," "Male, man or boy," "Nonbinary or

Table 2: Overview of the four toxicity detection systems. During training, all neural network models are trained on text, but only MuTOX-ASR and MuTOX are trained jointly with speech data. At inference time, only MuTOX has access to raw speech, while all other models rely on ASR text only.

Model	Type	Train		Inference	
		Text	Speech	ASR Text	Raw Speech
ETOX	Wordlist	-	-	✓	✗
DETOXIFY	Neural network	✓	✗	✓	✗
MuTOX-ASR	Neural network	✓	✓	✓	✗
MuTOX	Neural network	✓	✓	✓	✓

gender non-conforming," and "Transgender." Selectable groups were a superset of those used in Civil Comments (Borkan et al., 2019), though annotators could provide free-text responses when the provided categories were insufficient. See appendix D for the full list of groups annotated.

Transcript correction. After toxicity and group annotation, annotators were shown the audio's ASR transcript and asked "Does this transcript match the audio?" For the 21% of samples where the transcript was inaccurate, annotators were required to correct it manually.

Before Stage 1, annotators conducted a pilot analysis of 20 samples (later discarded) to evaluate the interface and identify issues with the guidelines. Annotators met frequently throughout Stage 1 to discuss problem cases and refine the guidelines, particularly regarding group annotation. In total, each annotator reviewed approximately 950 samples, spending approximately 30 to 45 seconds per sample. See table 1 for example annotations.

3.3 Stage 2: Individual review

Stage 1 responses were collated, and a majority vote was calculated for each sample. For questions allowing multiple selections (e.g., group mentions), the majority vote was the set of options selected by at least two annotators. Each annotator then independently reviewed the majority vote on a sample-by-sample basis. Annotators flagged samples where they disagreed with the majority vote for further discussion in Stage 3, alongside all samples where there was complete disagreement. Unflagged samples were assigned the majority vote as the final annotation.

3.4 Stage 3: Group review

Finally, annotators collectively reviewed all samples flagged during Stage 2, with the goal of sharing cultural knowledge and establishing consensus. Discussions were conducted in language-specific groups, where the annotator who flagged a sample presented their rationale, followed by a group discussion. Annotators were typically able to reach a consensus, but a "No consensus" label was occasionally assigned when annotators could not agree on a final label. Note that while "No consensus" indicates that the annotators cannot agree on an outcome, "Cannot say" indicates that annotators *agree* that toxicity could not be determined. For example, all annotators might concur that the sample's interpretation depends on external context, such as the identity of the speaker or audience. See fig. 1 for a summary of the final annotations.

4 The role of speech context

We compare four representative toxicity classifiers to evaluate the utility of using speech data directly as opposed to cascaded ASR-based systems, and to isolate the role of speech during training from during inference (see table 2).

4.1 Toxicity classifiers

ETOX (Costa-jussà et al., 2023) is a text-only wordlist-based classifier that supports 200 languages. While offering extensive coverage, it will only detect lexical toxicity and cannot account for context-dependent toxicity in polysemous words.

DETOXIFY "multilingual" (Hanu, 2020) is a text-only neural network that supports 7 languages and is trained on Wikipedia comments (Adams et al., 2017) and Civil Comments (Borkan et al., 2019), automatically translated using Google Translate.

MuTOX (Costa-jussà et al., 2024) is a multilingual neural network that supports 30 languages, trained on the MuTOX dataset. MuTOX is trained jointly on speech and text data encoded using SONAR (Duquenne et al., 2023). At inference time, it operates on both speech audio and an accompanying text transcript.

MuTOX-ASR is similar to MuTOX, but only

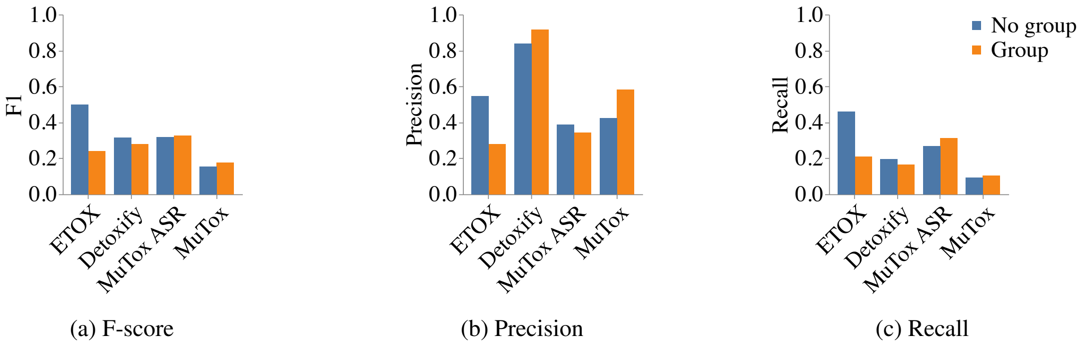


Figure 2: (a) F-score, (b) precision, and (c) recall of each classifier, for samples with and without group mentions. ETOX and DETOXIFY show lower F_1-score when a group is mentioned, whereas MUTOX-ASR and MUTOX show a slight increase. MUTOX is the only classifier to increase both precision and recall when groups are mentioned.

has access to SONAR text embeddings at inference time. MUTOX-ASR can only access ASR transcripts but may benefit from improved representations developed during joint training.

4.2 Methods

For each of the four models, we extract predictions for every sample in the English and Spanish MUTOX test sets. For ETOX, this is via lexical matching, whereas the model-based approaches all return a continuous toxicity score, subsequently binarized using a threshold. To ensure a fair comparison among all classifiers, the threshold was tuned on a per-language basis using the MUTOX validation partition to match the precision of ETOX. We evaluate each model's performance using F_1-score, precision, and recall, and evaluate their bias against group mentions using false positive rate (FPR), following Dixon et al. (2018).

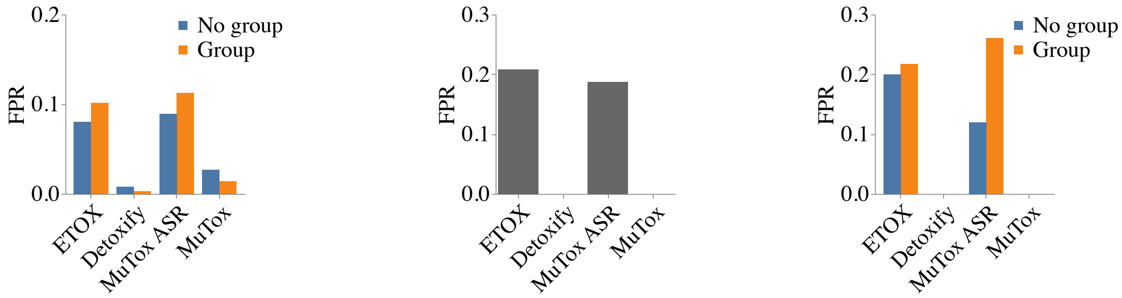
4.3 Results

Our evaluation reveals differences in the performance of speech-based and text-based toxicity detection models when sensitive groups are mentioned. Figure 2 shows that models relying solely on text (ETOX, DETOXIFY) exhibit a reduced F_1-score. On the other hand, both models trained with speech data (MUTOX-ASR, MUTOX) show a slight increase in F_1-score, but it is only the model with access to speech *at inference time* (MUTOX) that shows an increase across both precision and recall. Overall, while MUTOX shows the worst F_1-score of all classifiers, its precision is markedly higher than MUTOX-ASR (given equivalent threshold tuning), which is particularly important in reducing false positives.

Turning to FPR, fig. 3a shows clear differences between classifiers. Wordlist-based ETOX exhibits a high FPR that increases further when groups are mentioned, as does speech-trained MUTOX-ASR. In contrast, DETOXIFY and MUTOX both show low FPRs which decrease on group mentions. While the high FPR for ETOX is expected given the coarse nature of a wordlist, the differences between MUTOX-ASR (increase FPR on group mention) and MUTOX (decrease on group mention) are particularly interesting. Both models are trained jointly with speech and text data, but only MUTOX has access to speech data at inference time. This suggests that if a model is trained on both speech and text, then making speech unavailable at inference time worsens anti-group bias. This may be due to an over-reliance on group mentions as cues in the absence of important speech context.

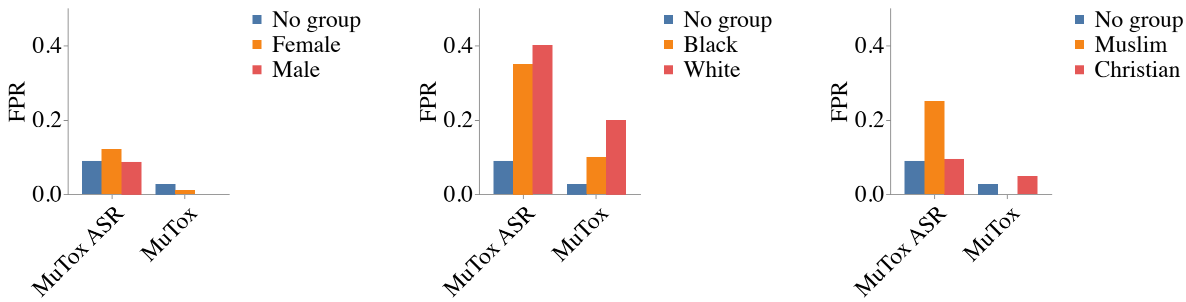
Ambiguous samples—those labeled "Cannot say" or "No consensus"—are a particular challenge for the wordlist-based ETOX and MUTOX-ASR, while DETOXIFY and MUTOX show an FPR of 0% (fig. 3b). Once again, we see an increase in FPR when groups are mentioned for MUTOX-ASR (fig. 3c). This also supports our hypothesis that models trained to process speech but unable to leverage speech at inference time struggle to separate group mentions from toxicity.

In fig. 4, we compare the FPR of MUTOX and MUTOX-ASR on specific group mentions to further isolate the effect of incorporating speech data during inference. With respect to gender (fig. 4a), MUTOX-ASR exhibits a higher FPR on samples mentioning women compared to samples with no group mentions, whereas MUTOX shows a reduced FPR when samples mention either women or men. Regarding race (fig. 4b), both models show a higher FPR for samples mentioning Black people



(a) All samples (b) Ambiguous samples (c) Ambiguous by group

Figure 3: (a) Classifier false positive rate (FPR) for samples with and without group mentions. (b) FPR of each classifier on samples annotators marked as "Cannot say" or "No consensus." (c) FPR on ambiguous samples with and without group mentions. DETOXIFY and MUTOX have an FPR of zero on ambiguous samples, while both ETOX and MUTOX-ASR demonstrate increased FPR when ambiguous samples mention groups.



(a) Gender identities (b) Racial or ethnic groups (c) Religious groups

Figure 4: False positive rate (FPR) of MUTOX and MUTOX-ASR on samples mentioning specific (a) gender identities, (b) racial or ethnic groups, (c) religious groups. (a) Mutox ASR shows a higher FPR for samples mentioning women than for other samples, whereas MUTOX's FPR decreases. (b) MUTOX-ASR shows a stronger bias against samples mentioning either White or Black people when compared to MUTOX. (c) Similarly, MUTOX-ASR shows a stronger bias against religious group mentions than MUTOX.

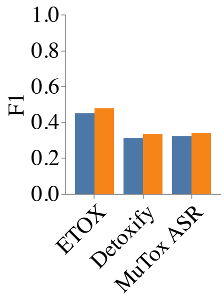
compared to no group mentions but unexpectedly show an even higher FPR for samples mentioning White people. As with gender, the increase in FPR for either group is reduced when incorporating speech during inference. For samples mentioning religious groups (fig. 4c), MUTOX-ASR shows a higher FPR for samples mentioning Muslims compared to samples mentioning no group, while MUTOX has an FPR of 0% on these samples.

Taken together, these results support our hypothesis that incorporating speech context during inference can help reduce toxicity detection failure and bias against certain groups, particularly for ambiguous or challenging samples. Notably, if a model is trained with speech data, our results suggest that it is important that the model *operates* on speech at inference time to avoid leveraging neutral group mentions as shortcuts for toxicity. That said, speech data is no panacea; speech-based models continue to exhibit biases in the form of increased FPR when certain groups are mentioned, suggesting systems should be deployed with caution.

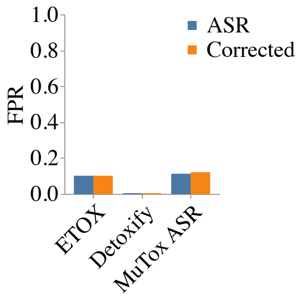
5 Effect of transcription error

One potential root cause of the failures observed in some cascaded ASR-based systems could be the ASR process. In other words, to what extent are the performance differences between the text-based classifiers a result of transcription failures rather than biases in the classifier itself? To address this question, we re-evaluate each classifier using the annotator-corrected transcripts.

In fig. 5a, we observe that correcting the transcripts leads to a predictable improvement in the overall performance of the text-based classifiers. At the same time, fig. 5b shows that the effect on the false positive rate (FPR) specifically for group mentions was minimal. This suggests that transcription errors alone do not account for the observed biases in toxicity detection when group mentions are present and that refining transcription pipelines



(a) English



(b) Spanish

Figure 5: (a) F_1-score of cascaded ASR-based classifiers with original ASR transcripts and annotator-corrected transcripts. (b) FPR on samples mentioning groups. Corrected transcripts only marginally improve model performance but have little to no impact on FPR.

is unlikely to be a productive strategy for reducing bias in speech toxicity detection systems.

6 Ambiguity in toxicity annotation

Our hypothesis that speech context can support less biased toxicity detection is predicated on the idea that toxicity itself is often highly subjective and context-dependent, making it hard to detect from outside of the initial conversation. Indeed, our annotation process is a testament to this fact. While we intentionally designed the annotation process with multiple stages to support interactive discussion and consensus building, an analysis of annotator disagreement demonstrates the extent to which toxicity judgments can vary.

After the first stage of annotation, annotators only unanimously agreed on toxicity in 66% of samples. For 32% of samples, at least two annotators agreed, producing a majority vote, but for the remaining 2% of samples, every annotator voted differently. A total of 7% of samples were flagged for Stage 2 discussion (see fig. 6a). These samples tended to be challenging to annotate, often requiring some degree of inference to determine what was left unsaid. After review, annotators could not agree (“No consensus”) on 8 samples, whereas 40 samples resulted in a “Cannot say” (see fig. 1a).

From the selection of flagged samples in table 3, we see that a variety of factors provoke discussion. For instance, annotators were unable to determine whether “you fuckers” (EN-3) was said in jest. The toxicity of EN-4 depends on whether “n***a” is pejorative or a re-appropriated word; annotators were instructed not to draw inferences about speaker identity. Sample ES-4 did not result in a consensus, as without further context, annotators were unable to determine the object referred to by “monstruo” (“monster”). Annotators were conflicted about whether EN-5 refers to the speaker’s viewpoint or to what others may say. While annotators leaned towards marking this sample as toxic, disambiguating between genuine toxicity on the part of the speaker and quotations or reading passages (e.g., ES-3) was a persistent challenge, even in the case of a recognizable Bible passage (ES-5).

Annotators also exhibited similar levels of disagreement when annotating for group mentions (see fig. 6b) despite our detailed and iterative shared guidelines. A particular challenge for annotators was identifying whether certain group mentions corresponded to the category “Racial or ethnic groups,” as speakers rarely disambiguate between nationalities, ethnicities, or linguistic groups. Ultimately, annotators reached a consensus after extensive discussion for all but one sample (see fig. 1b).

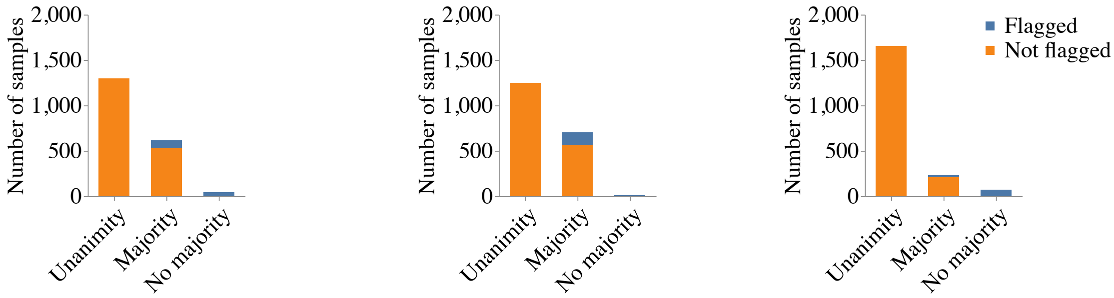
During transcription correction, annotators unanimously agreed more frequently—about 85% of samples, with only 5% requiring Stage 2 review. Annotators failed to reach a consensus on the correct transcription for 7 samples, highlighting the difficulty inherent in cascaded approaches.

7 Discussion & Conclusion

Leveraging our new, high-quality set of group annotations for the MuTox test partition, we compared the performance and biases of text- and speech-based toxicity classifiers. Our analysis revealed that models that make use of speech data during *both* training and inference exhibit reduced FPR bias against group mentions. For ambiguous samples, we found that models trained on speech but without speech access at inference time exhibit an increased FPR, suggesting that the multimodal models rely on spurious correlations when lacking an informative modality. Finally, we found that improving the quality of automated transcripts does little to reduce bias in English and Spanish, but this may change with lower-resourced languages where ASR systems exhibit poorer performance (Pratap et al., 2024).

7.1 The importance of multimodality

Speech is not simply spoken text—the two linguistic forms diverge in grammar, morphology, and register. As a richer medium (Daft et al., 1987), speech encodes more information that helps one



(a) Toxicity (b) Group mention (c) Transcription

Figure 6: Overview of inter-annotator agreement and review for (a) toxicity, (b) group annotation, and (c) transcription correction. Across all question types, annotators did not unanimously agree on a label for a sizeable proportion of the samples. Non-unanimous samples tended to have a majority vote, of which a reasonable fraction were flagged for Stage 2 review, alongside all samples lacking a majority.

Table 3: Selected ambiguous examples that provoked debate. Most samples lack sufficient context to fully understand the meaning of the segment, even with audio information. English translations in gray. MuTOX IDs in appendix E.

	Transcript	Toxic
EN-3	"We are gonna have a talk, you and me. You fuckers"	Cannot Say
EN-4	"He is your below average n***a. So what you need to do"	Cannot say
EN-5	"Yeah, you just gotta stop being gay and God will be okay with you"	Yes
ES-3	"Pero si yo lo venzo y lo mato, entonces seréis nuestros siervos" "But if I defeat him and kill him, then you will be our servants"	No consensus
ES-4	"…creó este monstruo en el Medio Oriente" "…created this monster in the Middle East"	Yes
ES-5	"Entreguen ahora a esos malvados de Gibea, para que los matemos y eliminemos así la maldad de Israel" "Now hand over those wicked people from Gibea, so that we can kill them and thus eliminate the wickedness of Israel"	Yes

better ascertain communicative intent. As such, even when the "words" converge, prosodic cues—e.g., inflection, tone, etc.—and contextual cues—e.g., speaker identity, social setting, etc.—in speech can contribute to differences in how meaning is construed in each of the two modalities (Kraut et al., 1992). By illustrating the improved performance of toxicity classifiers when speech data is introduced at inference time, we build on a growing body of work that demonstrates performance payoffs when engaging in multimodal and multitask learning.

7.2 Toxicity beyond social media

Much existing research on toxicity detection focuses on social media content moderation as the primary use case. As a result, toxicity detection datasets (e.g. Borkan et al., 2019) are often drawn from social media. This narrow focus may neglect increasingly relevant applications. For instance, with the general public's growing interaction with LLMs, it may be desirable to detect toxicity in generated responses, which may be orthogonal to determining whether the content itself is safe. Similarly, ensuring that machine translation systems do not introduce additional toxicity beyond what is present in the source is another emerging challenge (Sharou and Specia, 2022).

In contrast to earlier datasets, the MuTOX dataset is primarily extracted from "raw web corpora" (Costa-jussà et al., 2024, p. 2), representing a broader range of toxicity data. While this introduces certain biases (see §8), it reflects a positive shift toward evaluating toxicity in more diverse contexts beyond social media. As discussed in §6, it is *already* challenging for annotators to ascertain toxicity after the fact. As toxicity datasets expand to include novel application domains, new combinations of modalities (e.g. Kiela et al., 2020), and additional languages, robust annotation will become increasingly important.

7.3 Practical recommendations

To support the development of future speech toxicity datasets, we offer a few practical suggestions based on our experience annotating MuTOX.

Speech first. We recommend that annotators be instructed to focus principally on audio when evaluating speech toxicity. While audio may be unclear, ASR systems frequently make errors as they attempt to fill in gaps. During Stage 3 group review, many initially ambiguous samples became clearer when the original audio was considered.

Iterate and refine. Annotators should be encouraged to reference, discuss, and update a working set of annotation guidelines, particularly when dealing with edge cases. For example, while proper nouns were considered gender identity mentions, MUTOX's skew towards liturgical content (see §8) prompted extensive discussions about assigning gender to religious figures. Shared guidelines and regular discussion can improve annotation consistency, but there is rarely a single, definitive answer. When relying on crowd workers, where annotations are typically conducted in a single pass and disagreements resolved via majority vote, these nuances may be erroneously dismissed as noise.

Avoid automation. Recent work has explored using LLMs for annotation (e.g. Kumar et al. 2024) and benchmarking (e.g. Üstün et al. 2024). In inherently subjective and context-dependent tasks like toxicity detection, the majority of samples exhibit at least some form of ambiguity, with many samples requiring extensive discussion, consideration of possible interpretations, and understanding of historical and political context. Conducting annotation without human annotators in the loop is unlikely to adequately capture such intricacies.

8 Limitations

The MUTOX dataset comprises audio clips ranging from 2 to 8 seconds, often leading to truncated fragments. While annotators were instructed to make *small and reasonable* inferences when the truncated obvious was sufficiently predictable, the short clip length likely contributed to an inflated number of "Cannot say" responses. Truncated clips remove much-needed context (Pavlopoulos et al., 2020; Xenos et al., 2021), also amplifying the challenge of determining whether a speaker was expressing genuine toxicity or merely reading or quoting someone else. Recent work has suggested that models struggle to distinguish between counterspeech and harmful content (Gligoric et al., 2024), but our findings indicate that this issue also arises during the annotation process itself. Disambiguating between cases of "Cannot say" due to truncation versus genuine ambiguity would be more feasible with longer audio fragments, potentially improving annotation reliability.

Annotators were also discouraged from drawing inferences about speaker demographics, so annotators would typically assign a “Cannot say” to reappropriated words (e.g., in AAE). This approach may skew the distribution of toxicity labels for certain dialects. Annotators also observed a noticeable skew in the topic distribution across both the English and Spanish data, with several annotators remarking that a significant number of samples were fragments of Bible passages or religious sermons. Furthermore, the scope of our group annotations, intended for auditing rather than training, is limited by the time-intensive nature of annotation, with coverage constrained to English and Spanish. Future work should expand the sample size, domain diversity, and language coverage, particularly for under-resourced languages (Pratap et al., 2024), to better understand the broader impact of speech-based toxicity detection systems.

Due to the time-intensive nature of our iterative annotation process, this work only considers two languages, English and Spanish, both of which are higher-resourced and well-studied languages. Further research is required to understand the role of speech data in toxicity detection for lower-resourced languages. In particular, if ASR pipelines exhibit higher error rates in lower-resourced languages, then improving them might be a more productive strategy than our results for English and Spanish suggest.

Finally, we have not included statistical hypothesis testing when discussing our findings. Our principal contributions in this work are to produce a high-quality dataset of group annotations for multilingual speech toxicity detection, and subsequently to use those annotations to explore how classifier biases vary. As a result, we consider this work to be more exploratory than confirmatory (Bell and Kampman, 2021), and as such statistical hypothesis testing may not be appropriate .

References

CJ Adams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, Nithum Thain, and Will Cukierski. 2017. Toxic comment classification challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Çelebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and SEAMLESS Communication Team. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Samuel J. Bell and Onno P. Kampman. 2021. Perspectives on machine learning from psychology's reproducibility crisis. *Preprint*, arXiv:2104.08878.

Marcely Zanon Boito, Laurent Besacier, Natalia A. Tomashenko, and Yannick Estève. 2022. A study of gender impact in self-supervised models for speech-to-text systems. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022*, pages 1278–1282. ISCA.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 491–500. Association for Computing Machinery.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. 2022. Evaluating gender bias in speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2141–2147. European Language Resources Association.

Marta R. Costa-jussà, Mariano Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alexandre Mourachko, Christophe Ropers, and Carleigh Wood. 2024. MuTox: Universal multilingual audio-based toxicity dataset and zero-shot detector. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5725–5734. Association for Computational Linguistics.

Marta R. Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586. Association for Computational Linguistics.

Richard L Daft, Robert H Lengel, and Linda Klebe Trevino. 1987. Message equivocality, media selection, and manager performance: Implications for information systems. *MIS Quarterly*, pages 355–366.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35. Association for Computational Linguistics.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, 25(2):700–732.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 67–73. Association for Computing Machinery.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *Preprint*, arXiv:2103.15122.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s).

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. Gender representation in French broadcast corpora and its impact on ASR performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, AI4TV '19, pages 3–9. Association for Computing Machinery.

Sreyan Ghosh, Samden Lepcha, Sakshi Singh, Rajiv Ratn Shah, and Srinivasan Umesh. 2022. Detoxy: A large-scale multimodal dataset for toxicity classification in spoken utterances. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022*, pages 5185–5189. ISCA.

Kristina Gligoric, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959. Association for Computational Linguistics.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. In *Proceedings of the ACM on Human-Computer Interaction*, volume 6 (CSCW2). Association for Computing Machinery.

Vikram Gupta, Rini Sharon, Ramit Sawhney, and Debdoot Mukherjee. 2022. ADIMA: Abuse detection in multilingual audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pages 6172–6176. IEEE.

Laura Hanu. 2020. Detoxify. `https://github.com/unitaryai/detoxify`.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Robert Kraut, Jolene Galegher, Robert Fish, and Barbara Chalfonte. 1992. Task requirements and media choice in collaborative writing. *Human–Computer Interaction*, 7(4):375–407.

Shachi H. Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung-yi Lee, and Lama Nachman. 2024. Decoding biases: Automated methods and LLM judges for gender bias detection in language models. *Preprint*, arXiv:2408.03907.

Yi-Cheng Lin, Tzu-Quan Lin, Chih-Kai Yang, Ke-Han Lu, Wei-Chih Chen, Chun-Yi Kuan, and Hung-Yi Lee. 2024. Listen and speak fairly: A study on semantic gender bias in speech integrated large language models. In *IEEE Spoken Language Technology Workshop (SLT 2024)*, pages 439–446. IEEE.

Joseph Liu, Mahesh Kumar Nandwana, Janne Pylkkönen, Hannes Heikinheimo, and Morgan McGuire. 2024. Enhancing multilingual voice toxicity detection with speech-text alignment. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024*, pages 4298–4302. ISCA.

Yen Meng, Yi-Hui Chou, Andy T. Liu, and Hung-yi Lee. 2022. Don't speak too fast: The impact of data bias on self-supervised speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, pages 3258–3262. IEEE.

Mahesh Nandwana, Yifan He, Joseph Liu, Xiao Yu, Charles Shang, Eloi Du Bois, Morgan McGuire, and Kiran Bhat. 2024. Voice toxicity detection using multi-task learning. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024)*, pages 331–335. IEEE.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR.

Guilherme H. Resende, Luiz F. Nery, Fabrício Benevenuto, Savvas Zannettou, and Flavio Figueiredo. 2024. A comprehensive view of the biases of toxicity and sentiment analysis methods towards utterances with african american english expressions. *Preprint*, arXiv:2401.12720.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906. Association for Computational Linguistics.

Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180. European Association for Machine Translation.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. `https://github.com/HumanSignal/label-studio`.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939. Association for Computational Linguistics.

Alexandros Xenos, John Pavlopoulos, Ion Androutsopoulos, Lucas Dixon, Jeffrey Sorensen, and Leo Laugier. 2021. Toxicity detection can be sensitive to the conversational context. *Preprint*, arXiv:2111.10223.

A Annotator demographics

Samples were annotated by three annotators with native- or near-native-level proficiency in the sample's language. Near-native proficiency is to be understood as CEFR level C2. Annotators self-reported demographic information is described below.

Annotators had a mean age of 32.2 years with a standard deviation of 6.4 years. Three annotators described their gender as male, two as female, and one as nonbinary. Annotators described their ethnicity variously as White, White/Hispanic, Hispanic, or Middle Eastern, and were located in either France, the United Kingdom, or the United States. Among English annotators, two spoke American English and one spoke British English. For Spanish annotators, one spoke Cuban Spanish, one European Spanish and one Rioplatense Spanish.

B Annotation interface

See fig. 7 for the interface annotators used. Text transcripts were hidden until annotators had completed all other questions. See fig. 8 for the expanded transcription correction interface.

C Annotation guidelines

C.1 MuTOX toxicity guidelines

See Costa-jussà et al. (2024) for full details, but we include relevant sections here. MuTOX defines toxicity as "elements of language that are typically considered offensive, threatening, or harmful." Costa-jussà et al.'s definition spans:

- Profanities, defined as "language that is regarded as obscene, repulsive, or excessively vulgar, as well as scatological."
- Hate speech, defined as "language that is used to demean, disparage, belittle, or insult groups of people."
- Pornographic language, defined as "language that refers to sexual acts or refers in a vulgar way to body parts typically associated with sexuality."
- Physical violence or bullying language, defined as "language that is used to bully, threaten, silence individuals."

During iterative discussion, annotators agreed that *description* of violence, such as in a news report, should not be considered an example of the "physical violence of bullying language" category.

C.2 Group guidelines

We consider references to both groups as a whole and members of a group as mentions of that group. This includes implicit references, such as using a proper noun, gendered pronoun, or grammatical gender markers (except where the gender is the default, such as using a masculine gender marker to refer to mixed groups of people in Spanish). Annotators collectively constructed the examples in table 4 as a shared reference.

During iterative discussion, annotators also agreed on a number of small refinements to the annotation guidelines, which we include in the interest of transparency:

- References to *all* religions, *all* religious believers, etc., should not be considered mentions of a specific religious group.
- Annotating sexualities should be limited to mentions of sexual orientations, and exclude descriptions of sexual practices where a sexual orientation is not mentioned.
- While gender may sometimes be inferred from certain descriptions of sexualities, terms such as "gay" or "lesbian" should not be annotated as mentions of gender identities, unless gender is explicitly made apparent elsewhere in the sample.
- The use of the term "deformity" should be interpreted as a group mention of physical disability, unless otherwise indicated from the context.

D Group annotation results

After Stage 3 review, the following groups were assigned to at least one or more samples for each identity category.

Gender identities: "male, man or boy", "female, woman or girl", "transgender"

Sexualities: "homosexual, gay or lesbian", "queer", "bisexual", "heterosexual"

Religious groups: "christian", "jewish", "muslim"

Table 4: Examples of samples exhibiting group mentions presented to annotators for reference.

Group mention	Example	Reason
Yes	“All men suck”	Mentions men, as a group
Yes	“The woman drank her tea”	Mentions a specific woman
Yes	“She crossed the road”	Refers to a woman via a gendered pronoun
Yes	“Their pronouns are they/them”	Refers to a person using gender neutral pronouns
Yes	“The guy killed his houseplant”	Mentions a specific man
Yes	“As a queer person, I feel . . . ”	Mentions a specific queer person
Yes	“Jimmy’s best friend is gay”	Mentions a specific queer person
Yes	“Ella habla con su esposo”	Mentions a woman via a gendered pronoun and references a specific man
Yes	“Llámala, por favor”	Mentions a woman via a gendered pronoun
No	“They went to the party together”	Uses a gender-neutral plural pronoun
No	“Ustedes son los mejores”	Uses a gender-neutral plural pronoun
No	“Alex went shopping”	Gender neutral name

Table 5: Full MUTOX IDs for examples mentioned in tables 1 and 3.

Text ID	MUTOX ID
EN-1	0253d9e34712d19727de811a_2363424_2366142
EN-2	e7d68d1bcb15dd5ca0baa6d6_2394048_2396766
EN-3	25b2afe54ddab3f320478596_1324992_1329534
EN-4	e551701e4f0e2c64d58f4400_1536000_1540638
EN-5	106325b4a23644d7b5aad341_965280_968286
ES-1	255b54f0902d1919fbec7d86_5690208_5693598
ES-2	248e4f212b0ffe4ad99bc7d8_1683936_1687326
ES-3	5f575a5ab2945a3ffc6ab455_266400_271230
ES-4	e56d03f2445e80e9a864428b_352896_357150
ES-5	89def0fab1f03646e53c5589_376896_382974

Social classes or socio-economic statuses: “poverty”, “working class”, “agrarian”, “upper class”

Racial or ethnic groups: “white”, “african”, “afghan”, “russian”, “jewish”, “chinese”, “black”, “german”, “palestinian”, “english”, “french”, “indigenous american”, “irish”, “european”, “indian”, “ethiopian”, “arab”, “latino”, “egyptian”

E Example information

In the interest of brevity, samples mentioned in the main text are given a short identifier. See table 5 for the corresponding MUTOX IDs for all samples in tables 1 and 3.

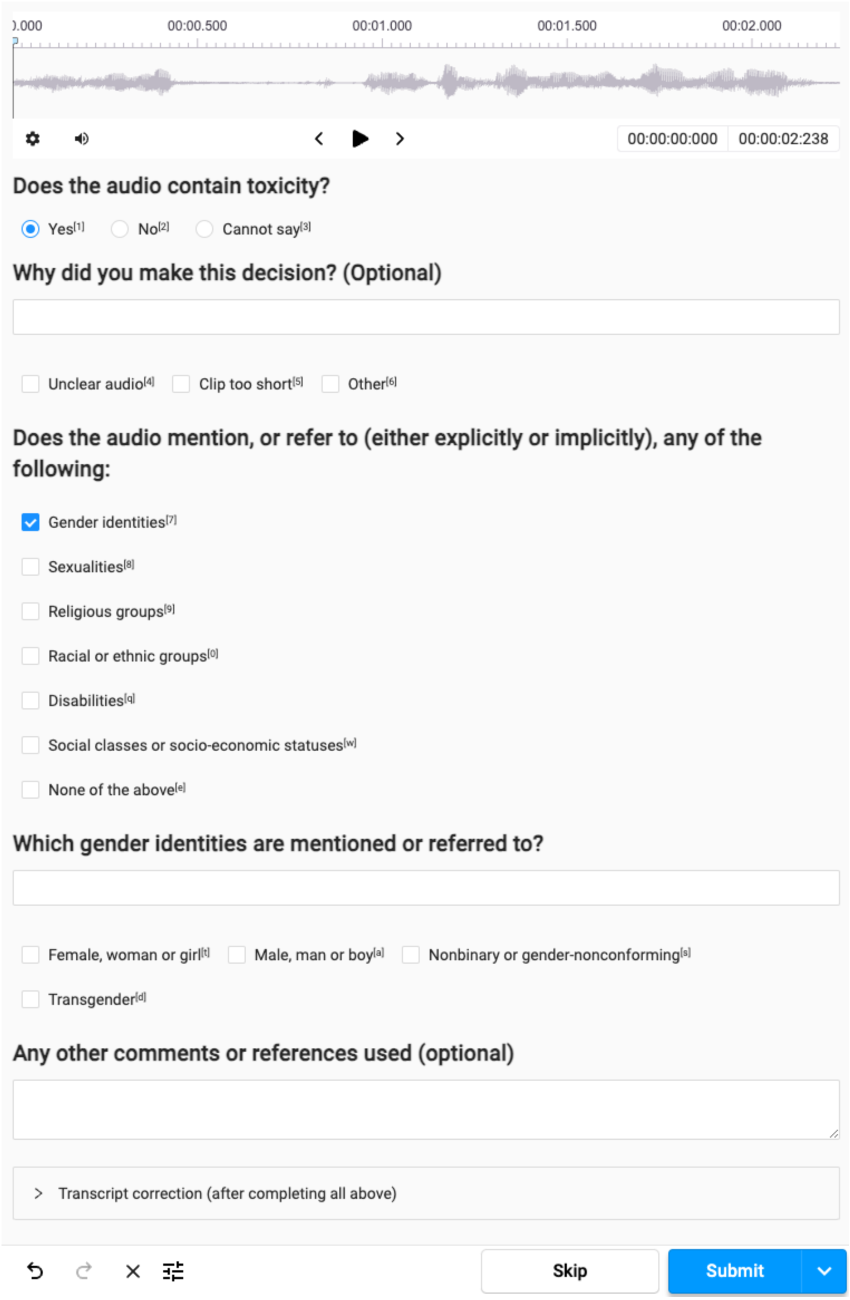


Figure 7: Annotation interface. Annotators could respond with free text if no checkbox was suitable.

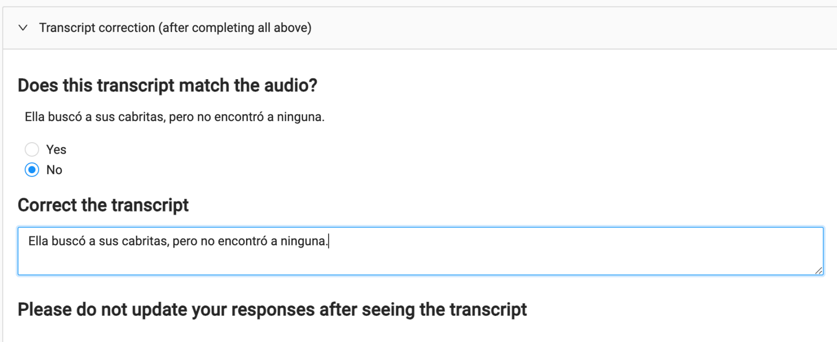


Figure 8: Transcription correction interface. Annotators were only asked to correct the transcript if they marked it as not matching the audio.