AUTALIC: A Dataset for Anti-AUTistic Ableist Language In Context

Anonymous ACL submission

Abstract

As our understanding of autism and ableism continues to increase, so does our understanding of ableist language towards autistic people. Such language poses a significant challenge in NLP research due to its subtle and context-dependent nature. Yet, detecting anti-autistic ableist language remains underexplored, with existing NLP tools often failing to capture its nuanced expressions. We present *AUTALIC*, the first benchmark dataset dedicated to the detection of anti-autistic ableist language in context, addressing a significant gap in the field. The dataset comprises 2,400 autism-related sentences collected from Reddit, accompanied by surrounding context, and is annotated by trained experts with backgrounds in neurodiversity. Our comprehensive evaluation reveals that current language models, including state-of-the-art LLMs, struggle to reliably identify anti-autistic ableism and align with human judgments, underscoring their limitations in this domain. We publicly release *AUTALIC* along with the individual annotations which serve as a valuable resource to researchers working on ableism, neurodiversity, and also studying disagreements in annotation tasks. This dataset serves as a crucial step towards developing more inclusive and context-aware NLP systems that better reflect diverse perspectives.

Trigger warning: this paper contains examples of anti-autistic ableist language including explicit slurs and references to violence.

1 Introduction

There are several critical frameworks used to define and understand disability, including autism (Lawson and Beckett, 2021). While these frameworks are ever-evolving, the medical model framework which defines disability as a "disease" is one of the most commonly applied in computer science research when discussing disability in general and autism specifically (Rizvi et al., 2024; Spiel et al., 2019a; Sideraki and Drigas, 2021; Anagnostopoulou et al., 2020; Parsons et al., 2020; Williams et al., 2023; Sum et al., 2022). This framework defines autism as a deficit of skills such as the Theory of Mind, and its applications in technology research largely focus on providing diagnosis and treatment to autistic people (Baron-Cohen, 1997; Begum et al., 2016; Rizvi et al., 2024; Spiel et al., 2019b). However, this framework has been criticized for creating a power imbalance between medical professionals and the autistic community as it promotes the belief that a formal diagnosis is necessary to validate the autistic identity (Bennett and Keyes, 2020). This belief also posits neurotypical behaviors as the "norm" and autism as a "deficit" of these norms, thereby promoting neuronormativity instead of neurodiversity, which views all neurotypes as valid forms of human diversity (Bottema-Beutel et al., 2021; Walker, 2014).

In contrast, prior work has shown viewing autism as an identity may increase the psychological well-being of autistic individuals and lower their social anxiety (Cooper et al., 2023). Although the language used to describe autism varies, prior studies with autistic American adults found 87% prefer identity-first language over person-first language (Taboas et al., 2023). Person-First Language (**PFL**) centers the person (i.e. person with autism), while Identity-First Language (**IFL**) centers the identity (i.e. autistic person) (Taboas et al., 2023).

Prior work examining ableist speech directed at content creators identified 11 types of "ableist hate and harassment" that they categorized as slurs and derogatory language, violent and eugenics-related speech, questioning ability and denying access, mocking and invalidation of disability identity, and objectifying the disabled body (Heung et al., 2024). Additionally, researchers have identified emerging terms such as "weaponized autism" that are gaining popularity on alt-right social media platforms, further reflecting the ever-evolving nature of nega-

tive societal attitudes towards autism (Welch et al., 2023).

Anti-autistic ableist language can be diverse in scope. It may include perpetuating stereotypes, using offensive language and slurs, or centering non-autistic people over the perspectives of autistic people (Bottema-Beutel et al., 2021; Rizvi et al., 2024; Darazsdi and Bialka, 2023). While abusive language detection systems can help identify such speech, they are known to demonstrate bias (Manerba and Tonelli, 2021; Venkit et al., 2022), with even LLMs perpetuating ableist biases (Gadiraju et al., 2023a). Additionally, anti-autistic ableist speech remains understudied, which is concerning given that classifiers trained on multiple hate speech datasets have shown a failure to generalize to target groups outside of the training corpus (Yoder et al., 2022).

Additionally, no datasets specifically focus on anti-autistic speech classification, and only 3 of the 23 datasets for bias evaluation in LLMs focus on disability (Gallegos et al., 2024). Prior work has also found that LLMs lack an acknowledgment of context, which leads to higher rates of false positives when classifying ableist speech (Phutane et al., 2024). These limitations are also found in toxicity classifiers, which excel primarily at identifying explicit ableist speech but may otherwise perpetuate harmful social biases leading to content suppression (Phutane et al., 2024).

Our work addresses these gaps in existing research by examining the performance of our dataset in the following domains for the task of anti-autistic hate speech classification:

1. The impact our dataset has on fine-tuning classifiers
2. The alignment of LLMs with the perspectives of trained human annotators
3. The extent to which in-context learning improves human-LLM alignment.

2 Related Work

Prior studies have also collected insights on anti-autistic language from the parents of autistic children (Allouch et al., 2018, 2019). However, toxic language datasets focusing on hate speech and abusive language have often addressed disability in general terms but have not explicitly focused on autism (ElSherief et al., 2018; Ousidhoum et al., 2019). Therefore, toxic language detection models, including LLM-based models, have been found to exhibit strong negative biases toward disabilities by classifying any disability-related text as toxic (Narayanan Venkit et al., 2023). Further, LLMs have been observed to perpetuate implicitly ableist stereotypes (Gadiraju et al., 2023b) and bias (Gama, 2024; Venkit et al., 2022). This, unfortunately, can sometimes be due to a research design that overlooks intra-community and disabled people's perspectives (Mondal et al., 2022), as well as autistic people's views, which may lead to harmful stereotypes (Rizvi et al., 2024; Spiel et al., 2019b). We make a step towards addressing these issues by building a dataset that focuses on ableist speech and autism by including autistic people's perspectives during the annotation processing as recommended by (Davani et al., 2023). *AUTALIC* contains all its labels and will also be useful for researchers interested in leveraging disagreements for difficult classification tasks (Leonardelli et al., 2021; Pavlick and Kwiatkowski, 2019).

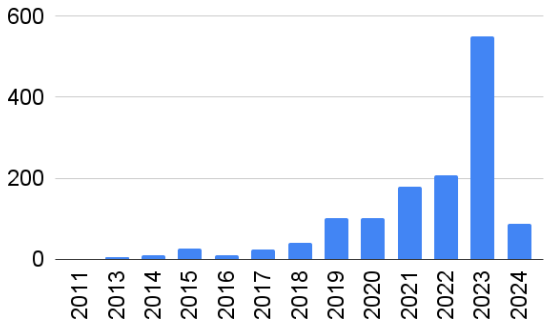

Figure 1: The distribution of years when *AUTALIC* posts were published on Reddit.

3 *AUTALIC*

This section discusses the process of compiling and annotating the *AUTALIC* dataset. We collected relevant sentences containing autism-related keywords from Reddit using the methods described in Section 3.1. The collected sentences were labeled by trained annotators, as discussed in Section 3.2.

3.1 Data Collection

Here, we present details from our data collection process, which include identifying our methodology for gathering representative sentences related to autism, curating the data obtained through our search parameters, and summarizing the resulting *AUTALIC* dataset.

Word	Sentence Count
autis*	1,221
ASD	226
disabilit*	184
aspergers	173
ABA	167
neurotyp*	158
aspie*	144
neurodiver*	103
AuDHD	99
disable*	93
autism speaks	66
stupid*	56
a11y	34
NT	27
retard*	25
idiot*	18
actually autistic	13
autism intensifies	6
ND	5
autie*	2

Table 1: The keywords included in our Reddit search and the number of sentences containing each term in the *AUTALIC* dataset.

3.1.1 Data Collection Criteria

First, we identified potential data sources and selected Reddit as the preferred source based on its popularity, focus on text-based content, and fewer API restrictions than X at the time of our data collection. We searched for keywords using the default search settings on Reddit, which filters posts based on relevancy by prioritizing rare words in the search query, the age of the post, and the amount of likes and comments it has. [1] Table 1 displays the keywords we searched for on Reddit and the final number of sentences with each keyword in our dataset.

We use the identified search terms detailed in Table 1 to collect two types of sentences:

- the target instance containing our keywords, to be labeled by the annotators,
- the sentences preceding or following target instances to provide additional context.

We collected 2,500 target sentences, with 2,014 preceding and 2,500 following them. Finally, we split our dataset into three parts by randomly selecting and assigning 800 unique target sentences to create three segments that were each annotated by a group of three annotators.

[1] https://support.reddithelp.com/hc/en-us/articles/19695706914196-What-filters-and-sorts-are-available

> Has anyone else shared their AuDHD Dx discovery and experienced an abject lack of curiosity and support

Figure 2: An example of a sentence from our dataset. The search keyword is shown in red, while the word in blue is an example of a word defined in our glossary.

3.1.2 Data Curation

As some of the identified keywords may appear in other contexts, we perform an exact word search for the acronyms to ensure unrelated words that might contain our acronyms are excluded from the search as they go beyond the scope of our dataset. For example, we searched for "applied behavioral analysis" and a case-sensitive search for *ABA*, which is a form of therapy intended to minimize autistic behaviors such as stimming (which is often used for self-soothing) (Sandoval-Norton et al., 2019). Similarly, we exclude any posts that are not written in English using the Python package `langdetect` [2] and posts that contain images, videos, or links.

3.1.3 Final Dataset

Our final dataset includes 2,400 sentences from 192 different subreddits. To protect our annotators' privacy, we have anonymized all labels.

Figure 1 shows the distribution of years in which the posts in *AUTALIC* were published. The range of publication years is 2013-2024, with most posts published in 2023. Figure 2 shows an example of a sentence from our dataset that uses both a search keyword and a word defined in our glossary described in Section 3.2.2.

The average number of likes on each post included in the *AUTALIC* dataset is 1,611.59, and only 135 posts have 0-5 likes. Table 2 details the subreddits from which we extracted the most significant number of sentences for our dataset.

[2] https://pypi.org/project/langdetect/

SubReddit	Sentence Count
r/Aspergers	116
r/Autism	88
r/AmITheAsshole	39
r/AutisminWomen	37
r/AuDHDWomen	24

Table 2: The subreddits with the most sentences included in the *AUTALIC* dataset and the number of sentences extracted from each.

Term	Definition
AuDHD	A combination of ADHD and autism
ND	Neurodivergent or neurodiverse
NT	Neurotypical

Table 3: Selected examples from our glossary detailing commonly used acronyms, abbreviations, and slang in autism discourse online. For the full glossary, see footnote [3].

3.2 Data Annotation

After obtaining sample sentences with contextual surrounding sentences, we proceeded to annotation: annotator selection, training, and data annotation.

3.2.1 Annotator Selection

We recruited nine annotators and randomly assigned them to annotate different segments of the dataset. Our annotators are undergraduate students who volunteered to annotate the data without financial compensation. They are culturally diverse and include people who grew up outside the United States. However, they are all fluent in English. Four of our annotators are gender minorities, and at least three self-identify as neurodivergent.

Although we ensured the annotators were from diverse backgrounds during our recruitment process, due to the collaborative nature of our annotation process, we cannot share the individual details of their identities. We also note that any personally identifiable information was destroyed upon the conclusion of our analysis and not shared outside of our research team.

3.2.2 Annotator Training

We provided a virtual orientation to all annotators explaining the history of anti-autistic ableism, examples of contemporary anti-autistic discrimination, and a brief overview of the annotation task.

The orientation begins with a discussion of the medical model approach to autism and its link to the Nazi eugenics program (Waltz, 2008; Sheffer, 2018). We define **neuronormativity** as the belief that the neurotypical brain is “normal” and other neurotypes are deficient in neurotypicality (Wise, 2023). We dive deeper into the medical model by discussing its impact on the self-perceptions and inclusion of autistic people in our society, such as an increase in suicidal ideation and social isolation among autistic people who mask or hide their autistic traits (Cassidy et al., 2014, 2018). Then, we cover the shifts in perspectives that emerged due to disability rights activism (Rowland, 2015; Cutler, 2019), and define **neurodiversity** as the belief that all neurotypes are valid forms of human diversity (Walker, 2014).

To explain the annotation task, we provide examples of sentences similar to what they may encounter while annotating. For example, we discuss how the inclusion of "at least" alters the connotations of the follow sentence:

> *At least* I am not autistic.

With just a minor change, the sentence can have an ableist connotation as it implies relief in knowing one is not autistic, as if it is shameful or wrong.

We also introduce our glossary to the annotators as a dynamic resource that can be altered as needed. This glossary contains words that may appear in autism discourse online that may not be commonly known to others. These include medical acronyms, slang, and references to organizations and resources commonly affiliated with the autistic community (such as Autism Speaks). Selected examples from our glossary are shown in Table 3. The final version of our glossary is available can be found in this footnote [3].

We conclude our orientation by providing a brief tutorial video demonstrating how to run the script that will guide each annotator through the annotation task.

3.2.3 Data Labeling

After completing the training, we assign each of the three segments of the dataset to three randomly selected annotators. Each annotator is assigned 800 unique sentences, with a goal of completing 200 annotations each week over four weeks.

Annotators select from three possible labels for each sentence: “Ableist,” “Not Ableist,” or “Needs More Context.” According to the Center for Disability Rights, “Ableism is a set of beliefs or practices

[3] https://figshare.com/s/3850e63295ccf6fceebb

Label	Definition	Count
-1	unrelated to autism or needs more context	595
0	not ableist	5,582
1	ableist	1,023

Table 4: An explanation of the labels used in our classification task and the resulting counts of each label from all annotators combined.

that devalue and discriminate against people with physical, intellectual, or psychiatric disabilities and often rests on the assumption that disabled people need to be 'fixed' in one form or the other."[4] We supplied this definition to annotators and instructed them to select the "ableist" label for any sentences that contain examples of these beliefs and practices directed towards autistic people.

Annotators are instructed to select the "not ableist" label for sentences that describe positive behaviors and attitudes regarding autism or disabilities or may be written by an autistic person reaching out for help and support. This includes individuals using medical terminology in a personal context (e.g., 'I need therapy'), discussion and suggestions from neurodivergent people (community-generated discussions), and general discussions of the medical processes (unrelated to neurodivergence). Some examples of statements labeled as not ableist are: "I am autistic", "As an autistic person, I think…"

The "Needs More Context" label is used for sentences an annotator is unable to definitively categorize as ableist or not ableist with only the surrounding sentences as context. This category includes text that is entirely unrelated to disabilities or remains ambiguous without additional context.

The labels used in our annotation task are -1 for sentences that are unrelated/need more context, 0 for sentences that are not anti-autistic, and 1 for sentences that are anti-autistic. The number of times our annotators assigned each label in our dataset is detailed in Table 4. To calculate our agreement scores, we consolidated labels -1 and 0 together based on feedback from our annotators that unrelated sentences needing more context could be classified as not anti-autistic in a purely granular classification.

While we use the majority label as the ground truth in our analysis, in our public dataset, we will be releasing the individual labels from each annotator due to a growing interest in embracing disagreements for such classification tasks in NLP (Leonardelli et al., 2021; Kralj Novak et al., 2022; Pavlick and Kwiatkowski, 2019).

[4]https://cdrnys.org/blog/uncategorized/ableism/

At this point people are watching us.

EM oh im so tired of people saying autism is a real disease.

Its from the vaccine.

Figure 3: The sentence in red in the middle is the target sentence that are our annotators are classifying as either ableist or not ableist. The sentences preceding and following this sentence are provided for more context.

3.2.4 Providing Context

While we provide additional sentences for context, the annotators are instructed to annotate the target sentence exclusively and only refer to the other sentences for additional context, such as determining whether the sentence is part of an intra-community discussion or the use of figurative speech (i.e. sarcasm). Figure 3 provides an example of a target sentence in context.

In this example, it is difficult to determine whether or not the writer had ableist intent, as it can be interpreted in multiple ways. For example, they can be critiquing the medical model, as many autistic activists do, thereby making it non-ableist. Or they could be genuinely promoting ableist misrepresentations. The contextual sentences help the annotators better understand the writer's intent. With these sentences, it is apparent that the writer is referring to the harmful and widely discredited association of vaccines with autism, which not only promotes anti-autistic ableism in society but also puts people's lives at risk by spreading disinformation about the benefits and harms of life-saving vaccines (Gabis et al., 2022; Taylor et al., 2014; Hotez, 2021).

Throughout the annotation process, annotators can edit previous annotations based on new knowledge to account for changes in language usage and connotations and the annotators' dynamic understanding of ableism.

It seems that the resistance I have been experiencing socially which clearly has only come from my own ways not even blaming HFA would only be intensified with this condition

Is anyone here a High Functioning Autist ?

Any input would be appreciated

Figure 4: An example of a sentence from our dataset with a high level of disagreement among annotators.

3.2.5 Disagreements

The Fleiss's Kappa scores were 0.125 for Team 1, 0.307 for Team 2, and 0.191 for team 3. These scores underline the difficulty of our classification task, which is apparent from the findings of prior work (Ousidhoum et al., 2019). We analyzed the sentences with the highest levels of disagreement in our dataset. We found that such posts:

1. tend to use medical model terminology or stereotypes
2. need additional context beyond the sentences we provided.

Our dataset contains 59 of these sentences.

Figure 4 contains an example of a sentence with a high disagreement among our annotators. While functioning labels are considered ableist due to their eugenicist approach of categorizing autistic people based on their perceived economic value (De Hooge, 2019), it is difficult to determine whether the original poster is autistic or not. The context is important here as classifying a sentence such as this as 'ableist' can lead to unfair censorship if the original poster is a self-diagnosed autistic person seeking advice. Therefore, these sentences were ultimately classified as 'not ableist' in *AUTALIC*.

Our analysis reveals a moderately strong negative correlation between the task completion time and agreement with other annotators that is statistically significant ($R = -0.644$, p-value: 0.0096). This highlights the importance of our orientation as we provided it simultaneously to the annotators, and the annotators who completed their task immediately after our orientation had higher agreement.

	Prec.	Recall	F1
LR	0.67	0.12	0.20

Table 5: The performance of Logistic Regression with Bag of Words.

4 Experiments

4.1 Data

Our dataset contains 2,400 sentences labeled as containing anti-autistic ableist language or not. The labels are obtained by calculating the mode from the three annotators of each data segment. Using this methodology, 242 target sentences contain examples of anti-autistic ableist language (10% of total), and 2160 sentences do not (90% of total).

4.2 Experimental Setup

We test the performance of different types of models on our dataset, including logistic regression, BERT, and LLMs ranging from PHI 3 Mini to GPT 4. These models are selected based on their diverse range of complexities. We also fine-tune BERT on our dataset to analyze the impact of fine-tuning on anti-autistic ableist speech classification.

4.2.1 Logistic Regression

We compare the performance of logistic regression (LR) supervised machine learning (ML) models to the majority of labels assigned by trained annotators. First, we use Bag of Words (BoW) as the features of the LR model. We run LR using 80%-20% train-test split of the entire dataset. These results are included in Table 5. We also compare model results with majority label downsampling since target sentences labeled with anti-autistic ableism only represent 10% of the dataset. We repeat the same LR analysis using Bag of N-Grams for both $n = 2$ and $n = 3$ (bi-grams and tri-grams).

4.2.2 BERT

We use the BERT (Devlin et al., 2019) large uncased model that contains 335M parameters, again using 80% of the data for training and 20% for testing. We used a linear layer on the pooled output of the model for classification. We obtained results with BERT after fine-tuning with the *AUTALIC* dataset, as shown in Table 6. We conducted this experiment using the test set as a validation set to choose the best hyperparameters, as we wanted to obtain the best-case-scenario performance for a fine-tuned BERT Large on this task.

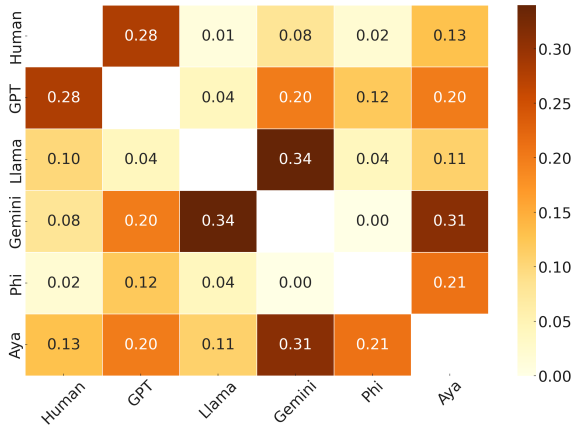

Figure 5: The Cohen's Kappa scores of each LLM comparing the agreement with human annotators and other LLMs after minor prompt engineering testing person-first or identity-first prompts.

4.2.3 LLMs

We prompt GPT-4o[5] and the instruction tuned variants of Llama-3-8B (Dubey et al., 2024), Aya-23-8b (Aryabumi et al., 2024), Gemma-2-2B (Team et al., 2024), Mistral-7B (Jiang et al., 2023), Phi-3-Mini (Abdin et al., 2024) to classify the sentences in our dataset, and we adjust the prompts to compare each LLM's performance using person-first or identity-first language. We keep the default parameters for each LLM to maintain consistency, and prompt them with the following questions:

1. Is this sentence ableist toward people with autism?
2. Is this sentence anti-autistic?
3. Is this sentence ableist toward autistic people?

We include each sentence from *AUTALIC* after the aforementioned questions in our full prompt. In addition, we provide preceding and following context for each target sentence to the LLM to mimic the level of the information supplied to human annotators.

4.3 Experimental Results

4.3.1 Fine-Tuning

Our experiments reveal that utilizing BERT for this classification task can lead to high rates of censorship. As BERT is not pre-trained with instructions unlike LLMs, we obtain its results after fine-tuning on *AUTALIC*. BERT has a high rate of false positives as nearly 2/3rds of sentences it considers ableist are not ableist according to our human annotators.

[5]https://openai.com/index/hello-gpt-4o/

BERT	Prec.	Recall	F1
Fine-Tuned	0.33	0.02	0.04

Table 6: The performance of BERT after fine-tuning on our dataset.

4.3.2 Human-LLM Alignment

Our assessment reveals that LLMs have low levels of alignment with human perspectives and the perspectives of other LLMs, which makes them unreliable agents for such classification tasks. We assess this alignment through a measurement using Cohen's Kappa scores. The scores shown in Figure 5 indicate that GPT-4o had higher levels of alignment with our human-annotated dataset than any other LLM ($k = 0.28$). Overall, the LLMs had low levels of agreement with human perspectives ($M = 0.1029, SD = 0.1096$), and even lower agreement with each other, as Phi had a worse-than-chance agreement with Gemma and Llama.

4.3.3 Safety Checks

Various LLMs have built-in safety checks, which cause them to give an invalid response, and this automatic censorship makes it difficult to employ them for the task of anti-autistic ableist speech classification. For example, Gemma has built-in safety checks that prevent it from labeling sentences that contain explicit slurs such as r*tard. Interestingly, Phi was the only LLM that did not classify certain sentences that mentioned video games without explicitly referencing autism. For example, posts discussing the Read Only Memories: NEURODIVER video game were included in our dataset as *neurodiver** was one of our search keywords. These sentences simply included the titles of video games and the dates, organizations, or people affiliated with them. Notably, these safety checks only occurred when the models were prompted to detect hate speech before classifying a specific sentence. This alludes to LLMs not having an understanding of anti-autistic language as hate speech without specific instruction.

4.3.4 Understanding Ableist vs. Anti-Autistic Speech

Our results with prompt engineering reveal that anti-autistic ableism is too abstract of a concept for LLMs to recognize, providing empirical evidence that their current reasoning abilities are not inclusive of the perspectives of autistic people. LLMs struggle with identifying anti-autistic speech re-

Model	Prec.	Recall	F1
GPT-4o	0.25	0.61	0.36
Llama3-8B	0.12	0.88	0.21
Aya-23-8B	0.15	0.66	0.24
Gemma-2-2B	0.11	0.91	0.22
Mistral-7B	0.17	0.65	0.24
Phi-3-Mini	0.15	0.43	0.22

Table 7: The mean performance of pre-trained LLMs with prompt engineering switching from PFL to IFL.

gardless of the terminology used, further indicating they are unreliable agents for data annotation tasks. Table 7 displays the results of prompt engineering using either person-first or identity-first language and using 'anti-autistic' as another identity-centered way of describing this form of ableism. These results show that LLMs struggle with conceptualizing ableism directed at autistic people and its semantics. However, we can see that GPT-4o performed considerably better than the other LLMs we tested. One key difference between GPT-4o and other LLMs is the number of parameters. Based on the performance of GPT-4o on benchmarks and our understanding of scaling laws in regards to model size (Kaplan et al., 2020), we estimate that GPT-4o has a far greater number of parameters than any other LLM we tested. This suggests that for larger model sizes, LLMs can begin to identify anti-autistic speech.

4.4 Discussion

Our experiments demonstrate the importance of *AUTALIC* in aligning LLM performance to human expectations in the contexts of autism inclusion and ableist speech classification. Through our experiments, we provide empirical evidence of the current limitations of using LLMs to identify expressions of anti-autistic ableism. These limitations include: a misalignment with human perspectives, a lack of understanding of the concept of anti-autistic ableism, and censorship arising from safety checks. Each of these limitations adds to the challenge of utilizing LLMs as reliable agents for such tasks.

5 Conclusion

In this paper, we introduced *AUTALIC*, the first benchmark dataset focused specifically on the detection of anti-autistic ableist language in context. Through the collection and annotation of 2,400 sentences from Reddit, we aim to fill a critical gap in current NLP research, which has largely overlooked the nuanced and context-dependent nature of ableist speech targeting autistic individuals. Our dataset not only captures key contextual elements but also incorporates a comprehensive annotation process led by trained annotators with an understanding of the autistic community, ensuring higher reliability and relevance.

Our experiments demonstrate the significant limitations of existing language models, including LLMs and traditional classifiers, in effectively identifying anti-autistic ableist speech. Standard pre-trained models such as showed poor performance, reinforcing the need for specialized fine-tuning. However, even after fine-tuning BERT on the *AUTALIC* dataset, our experiments reveal a high rate of false positives, which can lead to unfair censorship if BERT is employed for this task. Additionally, our results reveal that even state-of-the-art LLMs struggle with implicit bias and exhibit low agreement with human annotators on this task, further emphasizing the challenges of detecting subtle forms of ableism using generic models.

Looking forward, *AUTALIC* paves the way for significant advancements in content moderation systems, hate speech detection models, and research on ableism and neurodiversity. We envision this dataset as a cornerstone for future work in addressing bias against autistic individuals and fostering a more inclusive digital environment. By sharing this resource with the broader research community, we aim to catalyze the development of more equitable NLP systems that better serve underrepresented and marginalized groups.

6 Limitations

Despite doing our best to include a variety of ableist language against autistic people, our dataset can still show some selection bias (**?**) as we relied on keywords and specific social media threads to collect our data.

We acknowledge that our work centers Western perspectives in defining and understanding autism (Anderson-Chavarria, 2022). Therefore, we do not claim that *AUTALIC* is generalizable across languages and cultures as anti-autistic and ableist speech may manifest and be perceived differently.

Our dataset only contains sentences from Reddit, and subreddits focusing on autism such as r/Aspergers, r/Autism, and r/AutismInWomen were among the most popular subreddits the posts were

published in. Our search criteria included words like r*tard which is a slur broadly used against many intellectual disabilities and not specifically focused on autism. As well, since we broadened our search criteria to include variations of neurodiver* such as 'neurodiversity', 'neurodiverse', and 'neurodivergent', the search inevitably yielded posts discussing unrelated topics such as the NEURODIVER video game. The majority of these sentences were also published in 2023.

7 Ethics Statement

All the sentences in our dataset are publicly available, and we follow the methodologies of prior work in our data collection process (Atuhurra and Kamigaito, 2024). We received IRB approval from our university's review board and identified volunteer annotators through our association with various academic groups. Due to the sensitive nature of the content, we provided trigger warnings to our annotators and allowed them the freedom to annotate at their own pace. Additionally, we connected the members of each annotation team to enable discussions on the content and annotation process as needed.

While our dataset and citations will be made available to the academic community, commercial use of the dataset is not allowed due to the size and nature of the data. As our knowledge of anti-autistic ableism continues to evolve, *AUTALIC*'s classification might become outdated. While we will be adding disclaimers if needed to reflect these changes, we still encourage researchers building upon our work to stay updated on the latest semantics by referring to the perspectives of autistic scholars, activists, and organizations.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Merav Allouch, Amos Azaria, and Rina Azoulay. 2019. Detecting sentences that may be harmful to children with special needs. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1209–1213. IEEE.

Merav Allouch, Amos Azaria, Rina Azoulay, Ester Ben-Izchak, Moti Zwilling, and Ditza A Zachor. 2018. Automatic detection of insulting sentences in conversation. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pages 1–4. IEEE.

Panagiota Anagnostopoulou, Vasiliki Alexandropoulou, Georgia Lorentzou, Andriana Lykothanasi, Polyxeni Ntaountaki, and Athanasios Drigas. 2020. Artificial Intelligence in Autism Assessment. *International Journal of Emerging Technologies in Learning (iJET)*, 15(06):95.

Melissa Anderson-Chavarria. 2022. The autism predicament: models of autism and their impact on autistic identity. *Disability & Society*, 37(8):1321–1341.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Jesse Atuhurra and Hidetaka Kamigaito. 2024. Revealing trends in datasets from the 2022 acl and emnlp conferences. *arXiv preprint arXiv:2404.08666*.

Simon Baron-Cohen. 1997. *Mindblindness: An essay on autism and theory of mind*. MIT press.

Momotaz Begum, Richard W Serna, and Holly A Yanco. 2016. Are robots ready to deliver autism interventions? a comprehensive review. *International Journal of Social Robotics*, 8:157–181.

Cynthia L Bennett and Os Keyes. 2020. What is the point of fairness? disability, ai and the complexity of justice. *ACM SIGACCESS Accessibility and Computing*, 125(5):1–1.

Kristen Bottema-Beutel, Steven K Kapp, Jessica Nina Lester, Noah J Sasson, and Brittany N Hand. 2021. Avoiding ableist language: Suggestions for autism researchers. *Autism in adulthood*.

Sarah Cassidy, Louise Bradley, Rebecca Shaw, and Simon Baron-Cohen. 2018. Risk markers for suicidality in autistic adults. molecular autism, 9 (1), 42.

Sarah Cassidy, Paul Bradley, Janine Robinson, Carrie Allison, Meghan McHugh, and Simon Baron-Cohen. 2014. Suicidal ideation and suicide plans or attempts in adults with asperger's syndrome attending a specialist diagnostic clinic: a clinical cohort study. *The Lancet Psychiatry*, 1(2):142–147.

Kate Cooper, Ailsa J Russell, Jiedi Lei, and Laura GE Smith. 2023. The impact of a positive autism identity and autistic community solidarity on social anxiety and mental health in autistic young people. *Autism*, 27(3):848–857.

Emily Sheera Cutler. 2019. Listening to those with lived experience. *Critical psychiatry: Controversies and clinical implications*, pages 179–206.

Zoe Darazsdi and Christa S Bialka. 2023. "oh, you couldn't be autistic": Examining anti-autistic bias and self-esteem in the therapeutic alliance. *Autism*, 27(7):2124–2134.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Anna N De Hooge. 2019. Binary boys: autism, aspie supremacy and post/humanist normativity. *Disability Studies Quarterly*, 39(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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

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

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Lidia V Gabis, Odelia Leon Attia, Mia Goldman, Noy Barak, Paula Tefera, Shahar Shefer, Meirav Shaham, and Tally Lerman-Sagie. 2022. The myth of vaccination and autism spectrum. *European Journal of Paediatric Neurology*, 36:151–158.

Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023a. "i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 205–216, New York, NY, USA. Association for Computing Machinery.

Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023b. "i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 205–216, New York, NY, USA. Association for Computing Machinery.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Maysa Gama. 2024. Artificially intelligent and implicitly ableist: Analyzing language model bias towards persons with disabilities. *Accessible Canada - Accessible World / Un Canada accessible - Un monde accessible*.

Sharon Heung, Lucy Jiang, Shiri Azenkot, and Aditya Vashistha. 2024. "vulnerable, victimized, and objectified": Understanding ableist hate and harassment experienced by disabled content creators on social media. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Peter J Hotez. 2021. *Vaccines did not cause Rachel's autism: My journey as a vaccine scientist, pediatrician, and autism dad*. Johns Hopkins University Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. 2022. Handling disagreement in hate speech modelling. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 681–695. Springer.

Anna Lawson and Angharad E Beckett. 2021. The social and human rights models of disability: towards a complementarity thesis. *The International Journal of Human Rights*, 25(2):348–379.

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*.

Marta Marchiori Manerba and Sara Tonelli. 2021. Fine-grained fairness analysis of abusive language detection systems with checklist. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 81–91.

Ishani Mondal, Sukhnidh Kaur, Kalika Bali, Aditya Vashistha, and Manohar Swaminathan. 2022. "#DisabledOnIndianTwitter" : A dataset towards understanding the expression of people with disabilities on Indian Twitter. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 375–386, Online only. Association for Computational Linguistics.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2023. Automated ableism: An exploration of explicit disability biases in sentiment and toxicity analysis models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 26–34, Toronto, Canada. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Sarah Parsons, Nicola Yuill, Judith Good, and Mark Brosnan. 2020. 'Whose agenda? Who knows best? Whose voice?' Co-creating a technology research roadmap with autism stakeholders. *Disability & Society*, 35(2):201–234.

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

Mahika Phutane, Ananya Seelam, and Aditya Vashistha. 2024. How toxicity classifiers and large language models respond to ableism. *arXiv preprint arXiv:2410.03448*.

Naba Rizvi, William Wu, Mya Bolds, Raunak Mondal, Andrew Begel, and Imani N. S. Munyaka. 2024. Are robots ready to deliver autism inclusion?: A critical review. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Margaret Rowland. 2015. Angry and mad: A critical examination of identity politics, neurodiversity, and the mad pride movement. *Journal of Ethics in Mental Health*, 9:1–3.

Aileen Herlinda Sandoval-Norton, Gary Shkedy, and Dalia Shkedy. 2019. How much compliance is too much compliance: Is long-term aba therapy abuse? *Cogent Psychology*, 6(1):1641258.

Edith Sheffer. 2018. *Asperger's children: The origins of autism in Nazi Vienna*. WW Norton & Company.

Angeliki Sideraki and Athanasios Drigas. 2021. Artificial Intelligence (AI) in Autism. *Technium Social Sciences Journal*, 26:262–277.

Katta Spiel, Christopher Frauenberger, Os Keyes, and Geraldine Fitzpatrick. 2019a. Agency of Autistic Children in Technology Research—A Critical Literature Review. *ACM Transactions on Computer-Human Interaction*, 26(6):1–40.

Katta Spiel, Christopher Frauenberger, Os Keyes, and Geraldine Fitzpatrick. 2019b. Agency of autistic children in technology research—a critical literature review. *ACM Trans. Comput.-Hum. Interact.*, 26(6).

Cella M Sum, Rahaf Alharbi, Franchesca Spektor, Cynthia L Bennett, Christina N Harrington, Katta Spiel, and Rua Mae Williams. 2022. Dreaming Disability Justice in HCI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–5, New Orleans LA USA. ACM.

Amanda Taboas, Karla Doepke, and Corinne Zimmerman. 2023. Preferences for identity-first versus person-first language in a us sample of autism stakeholders. *Autism*, 27(2):565–570.

Luke E Taylor, Amy L Swerdfeger, and Guy D Eslick. 2014. Vaccines are not associated with autism: an evidence-based meta-analysis of case-control and cohort studies. *Vaccine*, 32(29):3623–3629.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nick Walker. 2014. Neurodiversity: Some basic terms & definitions.

Mitzi Waltz. 2008. Autism= death: The social and medical impact of a catastrophic medical model of autistic spectrum disorders. *Popular narrative media*, 1(1):13–24.

Christie Welch, Lili Senman, Rachel Loftin, Christian Picciolini, John Robison, Alexander Westphal, Barbara Perry, Jenny Nguyen, Patrick Jachyra, Suzanne Stevenson, et al. 2023. Understanding the use of the term "weaponized autism" in an alt-right social media platform. *Journal of Autism and Developmental Disorders*, 53(10):4035–4046.

Rua Mae Williams, Louanne Boyd, and Juan E. Gilbert. 2023. Counterventions: a reparative reflection on interventionist HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–11, Hamburg Germany. ACM.

Sonny Jane Wise. 2023. *We're All Neurodiverse: How to Build a Neurodiversity-Affirming Future and Challenge Neuronormativity*. Jessica Kingsley Publishers.

Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How hate speech varies by target identity: A computational analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.