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 dataset dedicated to the detection of anti-autistic ableist language in context, addressing a significant gap in the field. AUTALIC comprises 2,400 autism-related sentences collected from Reddit, accompanied by surrounding context, and annotated by trained experts with backgrounds in neurodiversity. Our comprehensive evaluation reveals that current language models, including state-of-the-art LLMs, struggle both to reliably identify anti-autistic ableism and to align with human judgments, underscoring their limitations in this domain. We publicly release AUTALIC along with the individual annotations. 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 ableist language including explicit slurs and references to violence.

1 Introduction

There are several critical frameworks used to define autism (Lawson and Beckett, 2021), including the medical model, which defines disability as a "disease" and is one of the most widely used in computer science research focusing on autism (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).

Since this framework defines autism as a deficit of skills, 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). 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., 2021a; Walker, 2014).

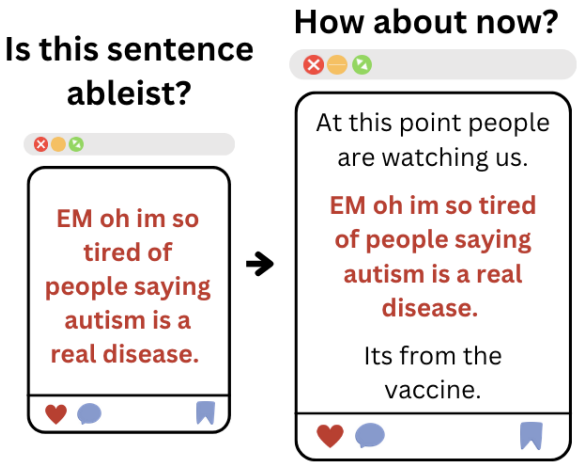

Figure 1: The example illustrates the importance of labeling sentences in context. The target sentence alone, shown on the left, is difficult to classify as ableist toward autistic people. Adding the surrounding sentences, as shown on the right, provides context revealing the original poster's reference to the debunked vaccine-autism stereotype, which is tied to anti-autistic stigma (Mann, 2019; Davidson, 2017).

To improve the alignment of AI research with neurodiversity, we present AUTALIC, a dataset of 2,400 autism-related sentences that we collect from various communities on Reddit, along with the original context of 2,014 immediately preceding and 2,400 following sentences. 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 final dataset contains all of the labels to capture the nuances in human perspectives, and allows AUTALIC to serve as a resource for researchers studying anti-autistic ableist speech, neurodiversity, or disagreements in general.

Through a series of experiments with classical models and 4 LLMs, we find empirical evidence highlighting the difficulty of this task, and that LLMs are not reliable agents for such annotations. Our evaluations indicate that reasoning LLMs have the most consistent scores regardless of the language used in the prompt, thereby indicating a more thorough understanding of the different ways anti-autistic speech may be identified or may manifest in text. We find that in-context learning examples provide mixed results in helping improve the task comprehension among LLMs.

2 Related Work

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., 2021a; 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).

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 (e.g. "person with autism"), while Identity-First Language (**IFL**) centers the identity (e.g. "autistic person") (Taboas et al., 2023). Supporting this finding, other researchers have found that viewing autism as an identity may increase the psychological well-being of autistic individuals and lower their social anxiety (Cooper et al., 2023).

Ableist language online varies, may manifest in different ways and is ever-evolving (Heung et al., 2024; Welch et al., 2023). 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). To our knowledge, there are no previous datasets specifically focused on anti-autistic speech classification, and only 3 of the 23 datasets for bias evaluation in LLMs focus on disability (Gallegos et al., 2024). LLMs may be limited in that they 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). 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).

3 AUTALIC

To build AUTALIC, 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.

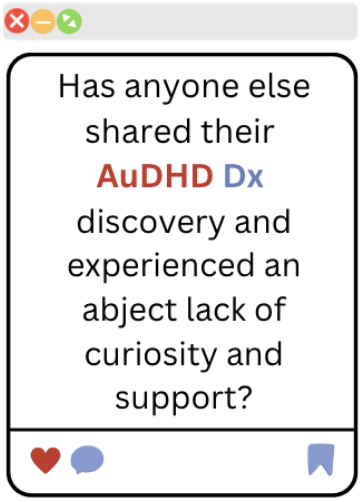

Figure 2: An example of a sentence from our dataset. The search keyword is shown in red, while the word in blue is an example of a word defined in our glossary.

3.1 Data Collection

We identify a methodology for curating sentences related to autism similar to prior datasets by collecting English-language sentences from Reddit (Overbay et al., 2023; Wadhwa et al., 2023; Njoo et al., 2023; Antoniak et al., 2024; Sabri et al., 2024; Allein and Moens, 2024; Buz et al., 2024). The limitations of this method are detailed in Section 6.

3.1.1 Data Collection Criteria

We select Reddit as our source based on its popularity, focus on text-based content, and fewer API restrictions than X at the time of our data collection in January 2024. We search for keywords using the default search settings, 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] The search terms include "autis*", "ASD", "aspergers", and "disabilit*"; the full list is available in the Appendix, Table 4.

We use the identified search terms to collect the target sentence instance containing our keywords, to be labeled by the annotators, and the sentences preceding or following target instances to provide additional context. We collect 2,400 target sentences, with 2,014 preceding and 2,400 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

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

Table 1: The subreddits with the most sentences included in the AUTALIC dataset and the number of sentences extracted from each.

The average number of likes on each post included in the AUTALIC dataset is 1,611.59. Table 1 details the subreddits from which the most significant number of sentences were extracted from. With the exception of r/AmITheAsshole, all of the other subreddits are autism-related.

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` and posts that contain images, videos, or links. [2]

3.1.3 Final Dataset

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

While nearly a quarter of the posts in our dataset were published in 2023, the range of publication years is 2013-2024. 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.

3.2 Data Annotation

3.2.1 Annotator Selection

We recruit nine upper level undergraduate researchers as volunteer annotators and randomly assign them to annotate different segments of the

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

dataset. We ensure that their involvement in our annotation process is voluntary, informed, and mutually beneficial. In particular, some participants choose to volunteer because they care deeply about the subject matter and wish to contribute to improving AI for autistic people. We select participants for whom this collaboration would provide relevant and valuable professional experience, grant them opportunities to engage in other aspects of the research, and provide mentorship and authorship recognition in accordance with the ACL guidelines.

We prioritize the well-being and autonomy of our annotators by providing full disclosure of the research process and subject material prior to their participation. We supply relevant trigger warnings, discuss the nature of the content in detail during an orientation session, and allow annotators to make an informed decision about whether they wish to proceed. We make them aware that they are free to withdraw from the study at any point.

Our annotators are US-based, culturally diverse, and include people who grew up outside the United States. They are all fluent in English. Four of our annotators are gender minorities, and at least three self-identify as neurodivergent. Although we ensure the annotators were from diverse backgrounds during our recruitment process, due to the collaborative nature of our annotation process, we do not 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 provide 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 following 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). An excerpt of our glossary is available in our Appendix. 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."

Ableist (1): We ask our annotators to select this label if a sentence contains ableist sentiments as defined by the Center for Disability Rights: "Ableism is a set of beliefs or practices 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."[3]

Not Ableist (0): Annotators select this label for sentences that describe positive or neutral behaviors and attitudes regarding autism, or posts 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

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

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

Table 2: An explanation of the labels used in our classification task and the resulting counts of each label from all 9 annotators combined.

therapy"), intra-community discussions, and general discussions of medical processes (unrelated to neurodivergence). Some examples of statements labeled as not ableist are: "I am autistic", "As an autistic person, I think..."

Needs More Context (-1): This label is used for sentences an annotator is unable to definitively categorize as ableist or not ableist even with the contextual sentences provided. This category includes text that is entirely unrelated to disabilities or remains ambiguous without additional context.

The number of times our annotators assign each label is detailed in Table 2. 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; Plank, 2022; Plank et al., 2014).

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).

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 1 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.

3.2.5 Disagreements

The average Fleiss's Kappa scores are 0.25. This score underlines the difficulty of our classification task, which is apparent from the findings of prior works (Ousidhoum et al., 2019), including a quantitative assessment of tag confusions that found the majority of disagreements are due to linguistically debatable cases rather than errors in annotation (Plank et al., 2014). Examples of such cases are provided in our Appendix.

We analyze the sentences with the highest levels of disagreement in our dataset. In 100 of these posts, we observe:

1. a tendency to use the medical model terminology or stereotypes (n=48)
2. a need for additional context beyond the sentences we provided

Figure 3 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"

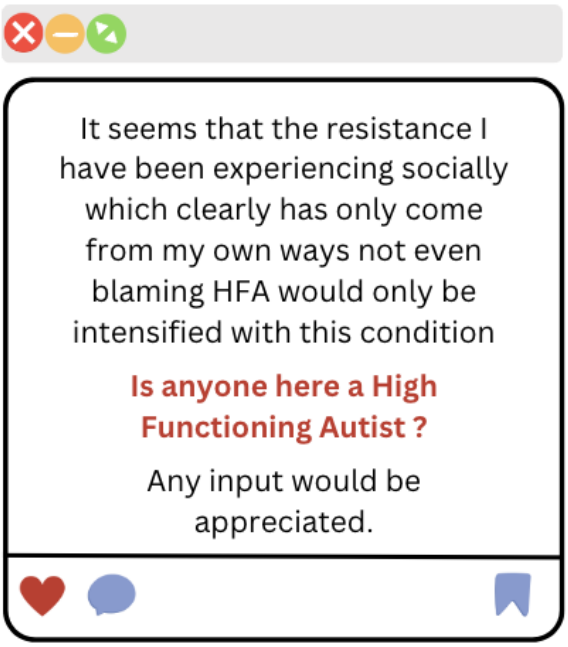

Figure 3: An example of a sentence from our dataset, shown in red, with a high level of disagreement among annotators.

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. The annotators who completed their task immediately after our orientation had higher agreement.

4 Experiments

4.1 Experimental Setup

We test the performance of different types of models on our dataset, including LLMs. These models are selected based on their diverse range of complexities. We also fine-tune BERT (Devlin, 2018) on our dataset to test classification.

4.1.1 Models

We use logistic regression (LR) with Bag of Words as the features, and using 80%-20% train-test split of the AUTALIC dataset.

Fine-tuned BERT We also utilize fine-tuned BERT (Devlin et al., 2019) as a baseline for LLM experiment results, using an 80%-20% train-test split. The F1 scores for LR and BERT (both pre-trained and fine-tuned) are presented in Table 3 alongside all LLM results.

4.1.2 Prompting LLMs

With our baseline established, we use the LLMs Gemma2 (Team et al., 2024), Mistral (Jiang et al., 2023), Llama3 (Dubey et al., 2024), and DeepSeek (Guo et al., 2025) to classify the sentences in our dataset, and we adjust the prompts to compare each LLM's performance. Due to limitations in computational resources, we are unable to fine-tune the LLMs.

Prompts Due to the ever-evolving nature of language and variations in preferences among autistic individuals (Taboas et al., 2023), we use three different kinds of prompts to measure the consistency LLMs have in their understanding of anti-autistic ableist speech. These prompts include person-first (i.e. 'people with autism'), identity-first (i.e. 'autistic people'), and conceptual (i.e. anti-autistic) language.

We keep the default parameters for each LLM to maintain consistency, and prompt them with the following questions:

> For each target sentence, respond to the following questions with 0 for no or 1 for yes. Refer to the preceding and following sentences if more context is needed.
>
> - Is this sentence ableist toward people with autism?
> - Is this sentence anti-autistic?
> - 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. We run two sets of experiments with each LLM: one that uses zero-shot prompting, and another containing engineered prompts for in-context learning verbatim from the definitions and examples provided in our annotator orientation (Appendix Section A.2).

4.2 Experimental Results

4.2.1 Fine-Tuning

Our experiments reveal that utilizing BERT for this classification task can lead to high rates of censorship. As BERT (unlike the other LLMs tested) is not pre-trained with instructions, we obtain its results after fine-tuning on AUTALIC. While the

Baselines			
Model	**Result**	**PT**	**FT**
LR	0.20	–	–
BERT	–	0.43	0.90
Simple Prompting			
LLM	**PFL**	**IFL**	**AA**
Gemma2	0.23	0.19	0.33
Mistral	0.28	0.27	**0.34**
Llama3	0.09	0.10	**0.15**
DeepSeek	0.58	0.57	**0.59**
In-Context Learning			
Gemma2	0.25	0.24	**0.34**
Mistral	0.31	0.24	**0.34**
Llama3	0.14	0.14	0.11
DeepSeek	0.55	0.56	0.55

Table 3: The F1 scores of various models using person-first (PFL), identify-first (IFL), and conceptual anti-autistic (AA) prompts with and without in-context learning examples for each LLM. The best scores for each model are in **bold**.

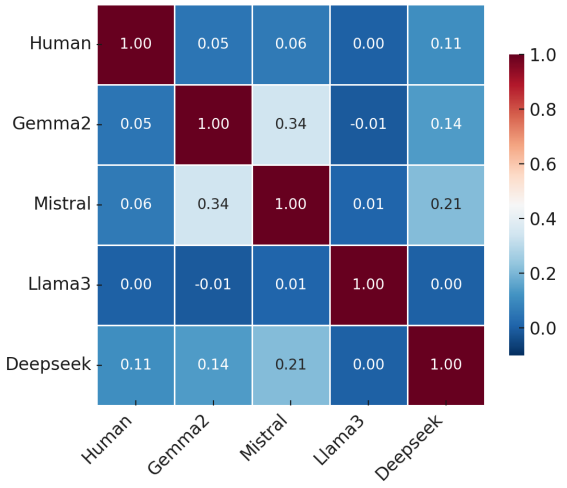

Figure 4: The mean Cohen's Kappa scores of each LLM comparing the agreement with human annotators and other LLMs.

pre-trained BERT model showed poor performance indicating that it was ineffective at identifying anti-autistic ableist speech, after fine-tuning on AUTALIC, the model's performance improved dramatically across all metrics, indicating that it performs better at predicting ableist speech correctly, has fewer false positives, and has a higher sensitivity to recognizing ableist speech.

4.2.2 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 4 indicate the highest level of alignment was demonstrated between Gemma2 and Mistral ($k = 0.34$). No LLM demonstrated alignment with our human-annotated dataset, although DeepSeek's alignment was notably higher than the others. Overall, the LLMs had low levels of agreement with human perspectives ($M = 0.091, SD = 0.110$). This indicates that LLMs with less than 10 billion parameters struggle with the task of classifying anti-autistic ableist language, even when provided with in-context examples.

4.2.3 In-Context Learning

After providing the in-context learning examples, Llama3 (+22.96%) and Gemma2 (+12.68%) display the biggest relative improvement in F-1 scores, indicating that both models benefit from the examples. In particular, we find our ICL examples resulted in large relative improvements in consistency of scores regardless of the language used in the prompts for each LLM. For example, providing Llama with examples helped decrease its relative change in F1 scores from 67.49% to 17.4% when switching from PFL to conceptual prompts, indicating a better understanding of the connection between anti-autistic ableism and ableism toward autistic people.

4.2.4 Understanding Ableist and 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 regardless of the terminology used, further indicating they are unreliable agents for data annotation tasks.

Table 3 displays the results of prompt engineering using either person-first or identity-first language and using "anti-autistic" to describe this form of ableism in more conceptual terms. Notably, switching from PFL or IFL to conceptual language in the prompts resulted in the largest relative changes in scores. For instance, prompting Llama

with "is this sentence anti-autistic?" instead of "is this sentence ableist toward people with autism?" resulted in a relative increase of 67.49%. These results show that LLMs struggle with understanding that "anti-autistic ableism" and "ableism toward autistic people" refer to the same phenomenon. Even after providing the ICL examples, the relative change in F1 scores between different prompts was as high as 32.66% for Gemma2.

Interestingly, DeepSeek, the only reasoning LLM we test in our study, has the best results and highest consistency out of all the other LLMs. Although its agreement with human annotators is low ($k : 0.11$), it is still double that of all other LLMs, as shown in Figure 4. This highlights the difficulty of this task, as more advanced reasoning is required to understand the nuances of anti-autistic ableism, including the terminology we use to describe such speech.

4.3 Discussion

Classifying anti-autistic ableist speech is challenging even within a Western, English-speaking context, since perceptions of what constitutes anti-autistic content differ significantly even among autistic individuals (Keating et al., 2023). Factors such as personal experiences, cultural norms, and evolving discourse around autism advocacy can influence each individual's perception of or sensitivity toward recognizing toxic speech, making it difficult to establish a consistent classification scheme (Ousidhoum et al., 2019; Bottema-Beutel et al., 2021b; Taboas et al., 2023; Kapp et al., 2013). While developing AUTALIC, we standardize our definition of anti-autistic ableist speech by providing our annotators with an orientation and a glossary. These resources are developed in alignment with the perspectives of autistic people (Bottema-Beutel et al., 2021a; Taboas et al., 2023; Kapp et al., 2013).

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 and traditional classifiers 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 a lack of agreement with each other even with in-context learning examples. Each of these limitations adds to the challenge of utilizing LLMs as reliable agents for such tasks.

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 exhibit low agreement with human annotators on this task, further emphasizing the challenges of detecting subtle forms of ableism using generic models. DeepSeek has the best performance out of all the LLMs in our study, further demonstrating the difficulty of this task, as it is the only a reasoning-focused LLM in our study.

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 and improve its alignment with neurodiversity.

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 (Ousidhoum et al., 2020) as we relied on keywords and specific social media threads to collect our data.

We recognize that, like other datasets presented at ACL, our work primarily reflects Western perspectives (Anderson-Chavarria, 2022; Kirk et al., 2023; Thapa et al., 2022; Aoyama et al., 2023; Takeshita et al., 2024; Park and Park, 2023). 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. However, we emphasize that AUTALIC is the first dataset to center the perspectives of autistic people in ableist speech detection, and move away from the tendency of AI to misclassify any text related to disability as 'toxic' (Narayanan Venkit et al., 2023; Heung et al., 2024; Van Dorpe et al., 2023; Hutchinson et al., 2020). Therefore, it can serve as a reference for future work exploring ableist speech in diverse contexts such as other cultures and languages.

Our dataset 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. While we did our best to filter out irrelevant posts, 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. While our dataset is small, the quality of our labels is high due to our standardized annotation protocols and resources that were developed using rigorous and iterative testing.

7 Ethics Statement

The data collected for our small-scale and non-commerical research is in compliance with Reddit's API limits and policies (Reddit, 2023a,b). 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 specifically recruit annotators for whom this collaboration and resulting paper authorship would be mutually beneficial, and provide a comprehensive overview of the task to ensure they make an informed decision to participate. Some of our annotators chose to volunteer as they care deeply about neurodiversity and autism inclusion.

We received IRB approval from our university's review board and identified volunteer annotators through our association with various academic groups. Given the sensitivity of the content, we provided annotators with appropriate trigger warnings, ensuring they could work at their own pace or withdraw from the study if necessary. 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. Refer to our Appendix for our Guidelines for Responsible Use.

References

Liesbeth Allein and Marie-Francine Moens. 2024. OrigamIM: A dataset of ambiguous sentence interpretations for social grounding and implicit language understanding. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 116–122, Torino, Italia. ELRA and ICCL.

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.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2024. Where do people tell stories online? story detection across online communities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7104–7130, Bangkok, Thailand. Association for Computational Linguistics.

Tatsuya Aoyama, Shabnam Behzad, Luke Gessler, Lauren Levine, Jessica Lin, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2023. Gentle: A genre-diverse multilayer challenge set for english nlp and linguistic evaluation. *arXiv preprint arXiv:2306.01966*.

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.

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

Kristen Bottema-Beutel, Steven K. Kapp, Jessica Nina Lester, Noah J. Sasson, and Brittany N. Hand. 2021b. Avoiding Ableist Language: Suggestions for Autism Researchers. *Autism in Adulthood*, 3(1):18–29.

Tolga Buz, Benjamin Frost, Nikola Genchev, Moritz Schneider, Lucie-Aim
'ee Kaffee, and Gerard de Melo. 2024. Investigating wit, creativity, and detectability of large language models in domain-specific writing style adaptation of Reddit's showerthoughts. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 291–307, Mexico City, Mexico. Association for Computational Linguistics.

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.

Michael Davidson. 2017. Vaccination as a cause of autism—myths and controversies. *Dialogues in clinical neuroscience*, 19(4):403–407.

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

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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, et al. 2024. The llama 3 herd of models. *arXiv 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.

Hanna Furfaro. 2018. New evidence ties Hans Asperger to Nazi eugenics program — thetransmitter.org. https://www.thetransmitter.org/spectrum/new-evidence-ties-hans-asperger-nazi-eugenics-program/.

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*.

John Greally. 2021. Autism glossary: Acronyms & abbreviations.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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.

Dieuwertje Dyi Huijg. 2020. Neuronormativity in theorising agency: An argument for a critical neurodiversity approach. In *Neurodiversity Studies*, pages 213–217. Routledge.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5491–5501.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Steven Kapp. 2019. How social deficit models exacerbate the medical model: Autism as case in point. *Autism Policy & Practice*, 2(1):3–28.

Steven K. Kapp, Kristen Gillespie-Lynch, Lauren E. Sherman, and Ted Hutman. 2013. Deficit, difference, or both? autism and neurodiversity. *Developmental Psychology*, 49(1):59.

Shanna Katz Kattari, EB Gross, Kari L Sherwood, and C Riley Hostetter. 2023. Infinity and rainbows. *Exploring Sexuality and Disability: A Guide for Human Service Professionals*.

Connor Tom Keating, Lydia Hickman, Joan Leung, Ruth Monk, Alicia Montgomery, Hannah Heath, and Sophie Sowden. 2023. Autism-related language preferences of english-speaking individuals across the globe: A mixed methods investigation. *Autism Research*, 16(2):406–428.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism. *arXiv preprint arXiv:2303.04222*.

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.

Benjamin W Mann. 2019. Autism narratives in media coverage of the mmr vaccine-autism controversy under a crip futurism framework. *Health Communication*.

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.

Michelle R Nario-Redmond, Alexia A Kemerling, and Arielle Silverman. 2019. Hostile, benevolent, and ambivalent ableism: Contemporary manifestations. *Journal of Social Issues*, 75(3):726–756.

Lucille Njoo, Chan Park, Octavia Stappart, Marvin Thielk, Yi Chu, and Yulia Tsvetkov. 2023. TalkUp: Paving the way for understanding empowering language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9334–9354, Singapore. Association for Computational Linguistics.

Disability Day of Mourning. 2024. Disability Day of Mourning; Remembering the Disabled Murdered by Caregivers. `https://disability-memorial.org/`.

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*.

Nedjma Ousidhoum, Yangqiu Song, and Dit-Yan Yeung. 2020. Comparative evaluation of label-agnostic selection bias in multilingual hate speech datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2532–2542.

Keighley Overbay, Jaewoo Ahn, Fatemeh Pesaran zadeh, Joonsuk Park, and Gunhee Kim. 2023. mRedditSum: A multimodal abstractive summarization dataset of Reddit threads with images. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4117–4132, Singapore. Association for Computational Linguistics.

Ashlee Owens. Adhd and autism: A paradoxical experience.

Min Su Park and Eunil Park. 2023. Acl ta-da: A dataset for text summarization and generation. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, page 1233–1239, New York, NY, USA. Association for Computing Machinery.

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*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Inc. Reddit. 2023a. Reddit data api terms. Accessed: February 9, 2025.

Inc. Reddit. 2023b. Reddit user agreement. Accessed: February 9, 2025.

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.

Adam Rosenblatt. 2022. Autism, advocacy organizations, and past injustice. *Brazilian Journal of Implantology and Health Sciences*, 4(6):450–467.

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.

Nazanin Sabri, Anh C. Pham, Ishita Kakkar, and Mai ElSherief. 2024. Inferring mental burnout discourse across Reddit communities. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 224–231, Miami, Florida, USA. Association for Computational Linguistics.

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.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Paolo Ponzetto. 2024. Aclsum: A new dataset for aspect-based summarization of scientific publications. *arXiv preprint arXiv:2403.05303*.

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, et al. 2024. Gemma 2:

Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Surendrabikram Thapa, Aditya Shah, Farhan Ahmad Jafri, Usman Naseem, and Imran Razzak. 2022. A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *CASE 2022-5th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, Proceedings of the Workshop*. Association for Computational Linguistics.

Autistic Union. 2012. Âû. https://www.facebook.com/AutisticUnion/posts/456270757745263/.

Josiane Van Dorpe, Zachary Yang, Nicolas Grenon-Godbout, and Grégoire Winterstein. 2023. Unveiling identity biases in toxicity detection: A game-focused dataset and reactivity analysis approach. In *Proceedings of the EMNLP 2023: Industry Track*, pages 263–274. Association for Computational Linguistics.

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.

Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. 2023. RedHOT: A corpus of annotated medical questions, experiences, and claims on social media. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 809–827, Dubrovnik, Croatia. Association for 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.

A Appendix

A.1 Guidelines for Responsible Use

A.1.1 Purpose and Scope

This dataset has been curated to aid in the classification and study of anti-autistic ableist language in a U.S. context using text from Reddit. It aims to support research and educational endeavors focused on understanding, identifying, and mitigating ableist speech directed at autistic individuals, while moving away from mis-classifying any speech related to autism or disability as toxic.

A.1.2 Applicability and Cultural Context

U.S.-Specific Results: The language examples and classification models in this dataset are primarily reflective of usage and cultural nuances in the United States. As a result, the dataset and any models developed from it may not be fully accurate or generalizable for other countries or cultural contexts.

Data Curation: The data included in this has been taken solely from posts and comments on Reddit and may not represent autism discourse on other platforms and in other contexts.

Contact for Latest Version: Language evolves over time. For the most up-to-date version of the dataset, or on more information on when the dataset was last updated, please contact the first author.

A.1.3 Access and Security

Password Protection: The dataset is password-protected to prevent unauthorized or automated scraping (e.g., by bots). While the password is publicly available as of this publication, it may require prior approval in the future as needed to ensure reponsible use.

Secure Storage: Users are expected to maintain secure protocols (e.g., encryption, controlled access) to prevent unauthorized sharing or leaks of the dataset. The dataset may not be shared without consent of the authors.

A.1.4 Permitted Uses

Free Use for Scientific Research: The dataset is publicly available without charge for legitimate scientific, academic, or educational research purposes, subject to the restrictions outlined below.

Academic and Non-Profit Contexts: Users in academic, research, or non-profit institutions may incorporate the dataset into studies, presentations, or scholarly articles, provided they follow these guidelines and appropriately cite the dataset and its authors.

A.1.5 Prohibited or Restricted Uses

Commercial Use:

Commercial use is not authorized without explicit written permission from the dataset authors. If you wish to incorporate the dataset into commercial products or services, you must obtain approval in advance.

Automated Content Moderation: Using the dataset to develop or deploy automated content moderation tools is not authorized without prior approval from the authors. This restriction helps ensure that any moderation system is deployed ethically and with proper considerations for context and language evolution.

A.1.6 Ethical Considerations and Privacy

Respect for Individuals and Communities: Users must handle the dataset with an understanding of the impacts of ableist language on autistic communities. The dataset's examples are provided solely for research and analysis and must not be used to perpetuate or normalize ableist attitudes, or to scrutinize or attack any individual annotators or original posters. This work is not intended as an ethical judgment or targeting of individuals, but rather an effort to improve AI alignment with the perspectives of autistic people.

Citations and Acknowledgements: When publishing findings, users should cite this dataset, acknowledging the work of its authors and the communities that provided the materials or data.

Compliance with Regulations: Researchers must comply with all relevant local, national, and international regulations and guidelines relating to data privacy and human subjects research where applicable.

A.1.7 How to Request Approval

Commercial or Moderation Use: If you intend to use the dataset for commercial purposes or automated content moderation, please submit a formal request, detailing:

- Project objectives
- Potential for data use and distribution
- Mechanisms to ensure ethical application and protection of the data
- The impact of the project, and its target end-users

Contact the First Author: All requests and inquiries should be directed to the first author, as listed in the dataset documentation or project website, available here: [REDACTED FOR PRIVACY]

A.1.8 Liability and Disclaimer

The dataset is provided "as is," without any guarantees regarding completeness, accuracy, or fitness for a particular purpose, especially outside of the U.S. context, for multi-media posts, or discussions on other platforms outside of Reddit.

User Responsibility: Users bear the responsibility for ensuring their use complies with these guidelines, as well as any applicable laws and ethical standards.

By accessing and using this dataset, you acknowledge that you have read and agreed to these Guidelines for Responsible Use, and that you understand the conditions under which the dataset may be utilized for your research or projects.

A.2 Annotator Orientation

A.2.1 Introduction

This subsection provides an overview of the annotation orientation session conducted for AUTALIC. The goal is to ensure annotators understand the history and contemporary examples of anti-autistic ableism, the importance of neurodiversity, and the different ways in which anti-autistic speech may manifest in text. Given the sensitive nature of this work, annotators are advised that they may encounter discussions involving ableist language, violence, self-harm, and suicide mentions.

A.2.2 Understanding Anti-Autistic Ableism

Anti-autistic ableism is the discrimination and devaluation of autistic individuals based on neuronormative standards. A striking example includes cases where caretakers harm autistic individuals due to societal stigma (of Mourning, 2024). The historical roots of such bias date back to Nazi-era eugenics research, where Hans Asperger categorized autistic individuals as either "useful" or "unfit," reinforcing a harmful hierarchical perception of autism (Furfaro, 2018).

Neuronormativity Neuronormativity is the societal belief that neurotypical cognition is the default and that neurodivergence is an abnormality (Huijg, 2020). This belief system marginalizes autistic individuals and contributes to discrimination in various aspects of life, including education, employment, and social interactions.

Deficit-Based Approaches and Their Harms Traditional medical models frame autism as a disorder requiring intervention or treatment (Kapp et al., 2013; Kapp, 2019). This perspective has led to:

- Increased exposure to violence and self-harm risk
- Social exclusion and stigmatization
- Internalized ableism and lower self-esteem

Benevolent Ableism Benevolent ableism refers to actions or attitudes that, while seemingly supportive, reinforce autistic individuals as "less than" neurotypicals (Nario-Redmond et al., 2019). Examples include organizations like *Autism Speaks*, which promote awareness campaigns that fail to center autistic voices (Rosenblatt, 2022). The use of symbols such as the puzzle piece is an example of this issue, as it implies that autism is a mystery to be solved rather than a valid identity.

A.2.3 The Neurodiversity Movement

The neurodiversity paradigm challenges the medical model by recognizing neurological variations as a natural and valid part of human diversity (Walker, 2014). Symbols such as the rainbow infinity sign inspired by the LGBTQ Pride flag have emerged from within the community to counter external narratives that frame autism as a deficit (Kattari et al., 2023).

Community Perspectives Autistic individuals often reclaim language and challenge neuronormative narratives. Important considerations for annotation include:

- Identity-first language (e.g., "autistic person" instead of "person with autism") is preferred by the majority of autistic adults in the United States (Taboas et al., 2023)
- Community-adopted terminology such as *Aspie* (a self-identifier used by some autistic individuals)

A.2.4 Annotation Tasks and Procedures

In this section, we provide an overview of the annotation task along with video examples of the process.

Common Annotation Challenges Annotators should exercise careful judgment when evaluating phrases. For example:

- Statements such as *"That's so autistic"* require contextual interpretation.
- The phrase *"This vaccine causes autism"* is categorized as ableist due to its history in promoting autism stigma.
- The subtle difference between *"I am not autistic"* and *"At least I am not autistic"* changes the meaning and must be carefully assessed.

A.2.5 Ethical Considerations and AI Bias

Challenges in Hate Speech Detection Research indicates that many existing AI models misclassify disability-related discourse as toxic, even when the content is neutral or positive (Narayanan Venkit et al., 2023; Venkit et al., 2022; Gadiraju et al., 2023a; Gama, 2024). Specific issues include:

- AI models exhibit over-sensitivity to disability-related discussions, frequently labeling them as harmful.
- AI models are more confident in detecting ableism when using *person-first language* (e.g., "ableist toward autistic people") than *identity-first language* (e.g., "anti-autistic"). *

**these are results from our preliminary study*

Project Overview This project seeks to mitigate biases in AI hate speech detection by:

- Training models using annotations informed by the neurodivergent community.
- Ensuring that AI does not misclassify community discourse as hate speech.
- Recognizing the distinction between hate speech and reclaimed terminology within the autistic community.

A.2.6 Resources

In this section, we provide resources such as our guidelines that contain a glossary to refer to or modify as needed. The terms in the glossary are those commonly used in neurodiversity discourse online.

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

A.2.7 Conclusion

The annotation orientation session is designed to equip annotators with the necessary knowledge to responsibly and accurately classify anti-autistic hate speech. By following the annotation guidelines and considering the broader socio-historical context, annotators contribute to the development of AI models that better serve neurodivergent individuals.

A.3 Search Keywords

This list of terms in Table 4 were used to identify target sentences on Reddit. The number of target sentences containing each term is included.

A.4 Glossary (Excerpt)

Below is an example of terms in our glossary. To view the full glossary, contact the first-author of this paper. This resource was created in collaboration with our annotators to define unfamiliar terms such as slang, medical abbreviations, laws, and other specialized language that they may encounter during the annotation process. In total, it contains 34 words.

Term	Definition
AuDHD	A combination of autism and ADHD (Owens).
Au, Âû	Used by autistic individuals to self-identify as autistic (Union, 2012).
ND	Neurodivergent, neurodiverse, or neurodiversity (Greally, 2021)

Table 5: Glossary of specialized terms used during annotation.

A.5 Challenging Cases

Here, we show examples of sentences with high disagreements among our annotators, along with their own notes on the aspects of each sentence that made it difficult to classify:

Example 1

> "ABA can be very beneficial to low-functioning autists who can't speak, use the bathroom, perform basic tasks, etc."

Promotes ABA and the "low-functioning" designation. **Example 2**

> "Autism can also lead to social issues like the autistic person saying the wrong thing at the wrong time with absolutely no utility involved."

Judgment that the autistic way is socially wrong.

Example 3

> "Autism or ADHD or otherwise, you still have to learn basic etiquette."

Specific to cultural context: Implies NT (neurotypical) etiquette is required, but no expectation for NT to learn ways of relating that are natural to autistic people.

Example 4

> "Basically, right after doing so much research, integrating with the autistic community, and accepting ASD as a part of myself, I was back to square one—left feeling like an idiot and immensely confused."

Slur against cognitive/intellectual disability, negativity associated with autistic identity, and medicalization of identity.

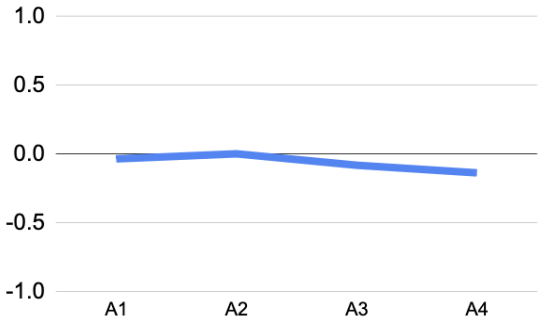

Figure 5: The self-agreement scores among annotators in a preliminary study highlight the difficulty of this task.

A.6 Self-Agreement Scores

In preliminary studies, we examined different labeling schemes for this task to identify the most efficient and effective annotation strategy. Our experiments revealed high levels of self-disagreement among annotators, as shown in Figure 5. The observed scores (M = -0.06, SD = 0.06) highlight the difficulty of the task and provide a meaningful baseline for comparison. Notably, our own annotation scores for AUTALIC were higher (M = 0.21, SD = 0.09), suggesting major improvement.

A.7 Annotation Platform

Figure 6 shows an example of an annotation task on our platform with contextual sentences.

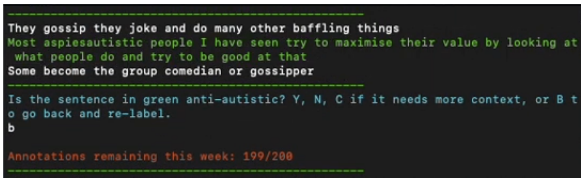

Figure 6: An example of an annotation task on our platform containing the target sentence (green) and contextual sentences (white).