

Common Sense or Ableism? Rethinking Commonsense Reasoning Through the Lens of Disability

Karina Halevy¹, Kimi Wenzel¹, Seyun Kim¹,
Kyle Bauer^{1,2}, Bruno Neira^{1,3}, Mona Diab¹, Maarten Sap^{1,4}

¹ Carnegie Mellon University, ² Lake Washington Institute of Technology,
³ Columbia University, ⁴ Allen Institute for Artificial Intelligence
khalevy [at] andrew [dot] cmu [dot] edu

Abstract

Commonsense reasoning (CSR) is a popular set of tasks used to evaluate large language model performance. However, what constitutes common sense might not apply to disabled people, leaving them out of AI progress by ignoring their identities. This paper investigates how datasets and models deal with disability in commonsense reasoning through a case study of three types of disabilities: autism, vision impairments, and long COVID. We first obtain annotations from disabled and non-disabled persons on $n = 300$ samples of three widely used CSR datasets (CommonsenseQA, Social IQA, Physical IQA) for *ableism*, which we define as entries in which gold labels assume a lack of disability, finding that $\sim 15\%$ of entries contain ableism. Then, we investigate LLMs' ability to flag ableism in CSR instances, and their flexibility to adapt to disability contexts—whether the CSR gold answer changes if the person is disabled. Our results show that the three LLMs tested have low sensitivity to human-detected ableism occurrences but still detect 5 to 25% of entries as ableist. Our findings call into question whether relying on such assumption-laden commonsense reasoning benchmarks risks leaving disabled people behind in the AI world.

1 Introduction

Commonsense reasoning (CSR) benchmarks are popular tools for measuring the progress of natural language processing models, including modern large language models (LLMs, Touvron et al., 2023a,b; Jiang et al., 2023; Abdin et al., 2024; OpenAI et al., 2024; The; Biderman et al., 2023; Team et al., 2025). However, what is considered *commonsense* is not always universal, and representing it as such may leave certain populations behind, such as the 1.3 billion disabled people worldwide¹ (Friedman and Owen, 2017; Cherney,

¹<https://www.who.int/news-room/fact-sheets/detail/disability-and-health>

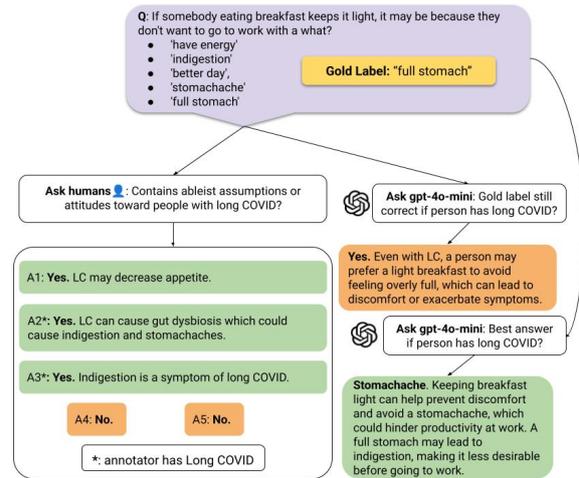


Figure 1: Commonsense QA (Talmor et al., 2019) entry. Annotators with Long COVID flag ableism. gpt-4o-mini also flags ableism when told that the person in the scenario has Long COVID. However, only one of three annotators without Long COVID flags ableism, and without the contextual information, gpt-4o-mini also does not flag ableism.

2011). For example (Fig. 1), “full stomach” is the gold label for a question in Commonsense QA (Talmor et al., 2019) that asks what a light breakfast before work would avoid. One of the 400+ million people worldwide with long COVID (LC) (Al-Aly et al., 2024) likely has gastrointestinal sequelae of COVID (Xu et al., 2023) that make “stomachaches” or “indigestion” more probable. While commonsense knowledge can be refuted when provided with extra context (Rudinger et al., 2020), considering some knowledge commonsense when it systematically ignores the experiences of specific groups of people can be harmful (Friedman and Owen, 2017; Cherney, 2011).

In this work, we conduct the first case study of how CSR data and models handle *disability*, focusing on long COVID,² vision impairment,

²We acknowledge that LC is not always considered a dis-

and autism, chosen based on the authors’ own disabilities. We choose these disabilities following disability studies and accessible computing field standards: research about disabled people should be conducted *with* these communities for appropriate service and representation (Stone and Priestley, 1996; Spiel et al., 2020). We study 900 entries across Commonsense QA (Talmor et al., 2019), Social IQA (Sap et al., 2019b), and Physical IQA (Bisk et al., 2020), all common CSR benchmark datasets for recent LLMs (Touvron et al., 2023b; Grattafiori et al., 2024; Abdin et al., 2024).

We first produce 5 sets of human annotations flagging *ableism* against each group, following the general definition of ableism put forth by Wolbring (2008) which defines it as “a set of beliefs, processes and practices that produce—based on abilities one exhibits or values—a particular understanding of oneself, one’s body and one’s relationship with others of humanity, other species and the environment, and includes how one is judged by others.” In our work, we operationalize ableism as the case in which the provided ground truth answer to a CSR question is significantly less likely to be correct if the person in the question is disabled (e.g., ‘indigestion’ becoming more likely than the gold ‘full stomach’ in Fig. 1).

With these annotations, we ask: **(RQ1)** To what extent do annotators agree on annotations of ableism? **(RQ2)** According to labels by annotators with each corresponding disability, how frequent is ableism in CSR datasets? Overall, human annotators find multiple entries with ableism, especially with CommonsenseQA (17.3% ableist according to disabled annotators, vs. 15.5% mean across datasets). Annotators also had relatively low agreement on ableism in each entry (e.g. Figure 1), but disabled annotators agreed more with each other than the full group.

We also stress test three LLMs’ capacity to reason about disability. We ask: **(RQ3)** How does LLMs’ flagging of ableism compare to human annotations? **(RQ4)** How well do LLMs *adapt* their CSR answers given information about disability? We find that all three LLMs struggle on both tasks, with balanced accuracies as low as 36% w.r.t. human gold labels on direct flagging and 34% on adapting.

Our findings suggest that lived experience

ability, but those with LC most often experience limitations that impair their daily activities, thus constituting a disability under the ADA. (HHS)

heavily influences one’s reasoning, that “common sense” is not a universal or objective concept, and that it is important to improve NLP benchmarks to better consider disabled identities.³

2 Related Work

Researchers have shown LLM generations to contain sexism, racism, Islamophobia, and ableism, among other biases (Wan et al., 2023; Kotek et al., 2023; Hemmatian et al., 2023; Abid et al., 2021; Hassan et al., 2021; Weidinger et al., 2022; Panda et al., 2025; Li et al., 2024a). Investigations on ableism have largely focused on toxicity prediction (Hutchinson et al., 2020) and text prediction (Hutchinson et al., 2020; Hassan et al., 2021). Within ableism, Gadiraju et al. (2023) find that LLM generations fixate on physical disability; violations of autonomy; sadness, loneliness, and the desire to be fixed; inspiration porn; and misinformation about disability—while they investigate model generations, we investigate similar phenomena in CSR benchmark data. On the data side, researchers have also developed benchmarks that have found LLMs to be subpar at detecting anti-autistic ableism in Reddit posts (Rizvi et al., 2025). Scholars have also introduced de-biasing tools (Esiobu et al., 2023) and advocated for the development of models and annotations alongside those affected by model biases (Chancellor et al., 2021; Gadiraju et al., 2023). However, despite CSR being an important task for tracking LLM progress, little work has examined ableism in CSR data. Appendix C provides more related work on ableism and the nuances of data annotation.

3 Problem & Concept Definition

We break down the task of quantifying ableism in CSR datasets into two problems, one for humans and one for LLMs. We examine both problems through commonly used datasets for LLM benchmarking (Touvron et al., 2023b; Jiang et al., 2023): **CommonsenseQA** (probing CSR about everyday scenarios; Talmor et al., 2019), **PIQA** (measuring CSR about physical activities; Bisk et al., 2020), and **Social IQA** (measuring CSR about the social implications of people’s actions; Sap et al., 2019b). Appendix D contains more details on the datasets.

Due to cost constraints and limited annotator availability, we randomly sample 300 items per dataset to annotate. We focus on three disabilities:

³Annotations will be publicized upon paper acceptance.

autism, long COVID (LC), and vision impairment, chosen based on the author team’s own lived experiences with disabilities.

4 Human Annotations

4.1 Methods

To address RQ1 and RQ2, the same five authors annotated each of our 900 entries. Annotators were asked only to perform these annotations on examples that: (a) involve exactly one person; and, (b) do not assume that the person is without disability (e.g. asking whether someone can “see” something assumes the person is not fully blind). The annotation procedure was as follows: For each entry, determine whether the question, context, answer choices, and correct answer taken as a whole are ableist against: (a) autistic people, (b) people with LC, and/or (c) vision-impaired people. In particular, in light of the fact that common sense itself inherently reflects a majority population opinion that may contain harmless variability, we ask annotators to flag ableism when an entry makes an assumption about a particular disability that would render the correct answer impossible/extremely unlikely, not just unpreferable but still plausible, for a person with that disability. Each annotator produces 3 binary annotations per entry, corresponding to the disabilities. Annotators also provide a free text justification for each answer that they marked as ableist. Annotators used both their lived experiences and the six questions for the annotation of data for disability content from [Gadiraju et al. \(2023\)](#) (Appendix C) to write rationales.

4.2 RQ1 Results: Annotator Agreement

Our first research question asks to what extent annotators agree on labels of ableism by examining inter-rater reliability ([McHugh, 2012](#)) via Cohen’s κ . Per Table 2, agreement is quite low, especially among non-disabled annotators and across disabled and non-disabled annotators. Agreement for vision impairment is lowest, even among vision-impaired annotators. This may be because the annotators’ vision impairments are corrected with eyeglasses in most situations. Thus, the remaining situations in which annotators would experience ableism would be highly activity-dependent and inconsistent—not all annotators partake in the same activities (e.g. swimming, dancing, chemical lab work) where regular eyeglasses are not allowed. In contrast, the annotators’ other disabilities are not correctable

Dis.	CSR Example	Rationale
LC	Q: Jordan rested every afternoon after eating a huge meal. How would Jordan feel afterwards? A: need food, easily exhausted, pretty full	If Jordan is resting after every afternoon meal they might have LC which makes them exhausted
Vision	Q: how do you show that you’re sad? A: put a smile on your face, put a frown on your face	Assumes emotions are communicated through the face
Autism	Q: How does applying for job while working impact someone? A: being employed, rejection, hunger, anxiety and fear, increased workload	Anxiety and fear could also be a part of it since you are probably anxious about the stability and enjoyability of your current job, neurodivergent people might have even more heightened emotions, especially when constructing multiple personalities to multiple employers

Table 1: Examples of flagged benchmark items (original correct answer in bold) and annotators’ rationales.

Disability	Dataset	All	Dis	Non-Dis
Autism	PIQA	0.05	-0.02	0.00
	SIQA	0.03	0.04	-0.02
	CQA	0.06	0.29	0.01
LC	PIQA	0.02	0.06	0.0
	SIQA	0.16	0.20	0.08
	CQA	0.09	0.20	0.11
Vision	PIQA	0.01	-0.01	-0.01
	SIQA	-0.01	-0.06	-0.02
	CQA	0.04	-0.03	0.00

Table 2: Inter-rater reliability (mean of pairwise Cohen’s κ s) for human-annotations. Bolded entries in each column indicate the highest agreement. *Dis* refers to the group of annotators with the specified disability; *Non-Dis* refers to those without.

and are more limiting for a more common set of daily activities (e.g. all annotators must eat and participate in social interactions), leading to more consistent experiences of ableism.

For autism and LC, we see the expected result that the annotators with the corresponding disability agree more with each other. Motivated by this result, **we consider the aggregated annotations of disabled annotators as the ground truth** for our following research questions.

4.3 RQ2 Results: Frequency of Ableism

To answer RQ2 (frequency of ableism according to disabled annotators), we calculated the number of instances where the disabled annotators considered the entry as ableist, aggregating the annotations both leniently (at least one annotator flagged

Dataset	Autism	LC	Vision	Total
PIQA	2.7 / 0.0	6.3 / 0.3	7.3 / 0.0	14 / 0.3
SIQA	10.3 / 1.3	8.3 / 1.7	2.7 / 0.0	15.3 / 0.3
CQA	8.7 / 2.0	9.7 / 1.7	2.0 / 0.0	17.3 / 0.3

Table 3: Table indicating percentages of entries found to be ableist by disabled annotators for each disability. Each cell is written x/y , where x is the % of entries that *at least one* annotator found ableist, and y is the % of entries that *both* annotators found ableist.

the entry as ableist) and stringently (both annotators flagged the entry). Per Table 3, annotators flagged approximately 14% of PIQA entries, 15% of SIQA entries, and 17% of CommonsenseQA entries as ableist in some form. CommonsenseQA has the highest frequency of LC-related ableism, PIQA’s ableism is roughly evenly split between LC and vision, and SIQA’s ableism is roughly evenly split between autism and LC. This makes sense, as SIQA is focused on social interactions, which are heavily determined by one’s neurotype. PIQA’s vision-related ableism likely relate to how physical activities can be heavily visually driven.

It is also unsurprising that all datasets contain at least 5% of entries with ableism against LC, as their creation predates the mass emergence of post-viral illness. The most common justifications for ableism in annotators’ free-text rationales were (1) that the data made an assumption about an agent’s integrity based on ability or (2) that the answer made an unnecessary assumption about ability. Appendix E.2 provides examples of these rationales.

5 LLM Experiments

5.1 Methods

We explore two LLM setups. *First*, to compare human and LLM annotations, we asked LLMs to flag ableism in the 900 dataset entries (RQ3), i.e. instances where the original answer is no longer correct given that the person is disabled. *Second*, we measure model adaptability to disability context in the CSR task (RQ4), inspired by defeasible reasoning (Madaan et al., 2021; Rudinger et al., 2020). We state that the person in the entry has one of our three disabilities of focus, then ask the model to select and justify the correct answer, adding a “none of the above” option. Appendix A provides the prompts. A positive ableism label is when the model answers differently than the dataset’s gold label. The ground truth for comparison is the annotations aggregated leniently among disabled

annotators (i.e. a positive label is when at least one disabled annotator flags ableism). The performance measure for both flagging and adaptability is balanced accuracy against this ground truth. We conduct this analysis for each of the three disabilities and examine three models: gpt-4o-mini⁴, gpt-5,⁵ and llama-3.3-70B-instruct-turbo.⁶ Table 4 shows balanced accuracies of each model’s annotations. An author with all three disabilities wrote both prompt templates, hence representing a realistic case of how a disabled user would interact with these models.

5.2 RQ3: LLM vs. Human Ableism Flagging

The “Flag (RQ3)” columns in Table 4 answer RQ3: the balanced accuracies for flagging ableism are low. The highest BA in this task is 59% (vision, PIQA, gpt-5). The main factor contributing to low balanced accuracies is the false negatives—when annotators flag ableism but the LLM fails to. Appendix E.1 shows recall and true negative rates. These results suggest that LLMs struggle to identify ableism in default answers within a benchmark.

5.3 RQ4 Results: LLM Adaptability

The “Adapt (RQ4)” columns in Table 4 show the balanced accuracies of all three models on the adapting task, addressing RQ4. Balanced accuracies are higher than for the flagging task in 15 of 27 cases but still low overall, with the highest at 70% (Social IQA, autism, llama-3.3-70B-turbo). As in RQ3, the main factor that contributes to low balanced accuracies is the false negatives. These results suggest that LLMs struggle to adjust their reasoning in line with human conceptions of ableism.

6 Discussion & Conclusion

This work conducts the first investigation of ableism in CSR datasets through human and LLM annotations on Physical IQA, Social IQA, and CommonsenseQA, finding ableism in a nontrivial proportion of entries ($\sim 15.5\%$). We also find that human annotators have low agreement on whether the entries are ableist, but those with lived experiences of the disabilities of focus tend to have higher agreement, aligning with prior research (Sap et al., 2019a; Al Kuwatly et al., 2020; Luo et al., 2020; Ross et al., 2017; Harding, 2013; Pohlhaus,

⁴<https://platform.openai.com/docs/models/gpt-4o-mini>

⁵<https://openai.com/gpt-5/>

⁶<https://www.together.ai/models/llama-3-3-70b>

Dataset	Disability	GPT-4o-mini		Llama-3.3-70B-Turbo		GPT-5	
		Flag (RQ3)	Adapt (RQ4)	Flag (RQ3)	Adapt (RQ4)	Flag (RQ3)	Adapt (RQ4)
PIQA	Vision	0.36	0.59	0.39	0.49	0.59	0.36
	Autism	0.51	0.34	0.51	0.40	0.51	0.34
	LC	0.50	0.50	0.50	0.55	0.50	0.50
SIQA	Vision	0.48	0.68	0.47	0.52	0.41	0.48
	Autism	0.48	0.43	0.44	0.70	0.49	0.55
	LC	0.46	0.58	0.39	0.45	0.50	0.41
CQA	Vision	0.45	0.39	0.43	0.44	0.45	0.42
	Autism	0.51	0.53	0.48	0.61	0.53	0.45
	LC	0.45	0.52	0.49	0.51	0.51	0.47

Table 4: Balanced Accuracy (BA) across all three models for the Flagging (RQ3) and Adapting (RQ4) tasks. Bolded values represent the highest performance for that specific model/task combination across disabilities. All models show low sensitivity to disability context across datasets.

2002). Finally, we find that both LLMs are insufficiently sensitive to disability contexts. Our findings point to two main needs for future work. First, CSR data sets and LLM benchmarks must be updated to include lived experiences of people with disabilities. Second, future work in developing LLMs should focus on improving their ability to change their reasoning patterns given disability information. With new developments in LLMs’ abilities to ask follow-up questions when they do not have enough information provide definitive responses upfront (Li et al., 2024b), work in these two directions may steer the LLM development and evaluation community away from commonsense reasoning as an objective benchmark altogether.

Limitations

We investigated three different datasets: SocialIQA, PIQA, and Commonsense QA. Although these three datasets were chosen to investigate data that represent everyday scenarios in three different perspectives, future work can expand our investigation to include more datasets, both from the list of datasets on which LLaMa-2 was evaluated (Touvron et al., 2023b) and from other LLMs’ reported benchmark datasets (Abdin et al., 2024; OpenAI et al., 2024). Examples include:

- HellaSwag (Zellers et al., 2019): 39.9k labelled training entries, 10k labelled validation entries, and 10k unlabelled test entries for commonsense natural language inference. Task formulation: given a sentence s and three more sentences as choices for logical follow-ups to s , pick the one correct follow-up.
- WinoGrande (Sakaguchi et al., 2019): a dataset for common sense question answer-

ing, for which there are several versions, one of which is the “debiasedversion”, said to reduce spurious correlations between instances and labels. This version has 9.25k train entries, 1.27k validation entries, and 1.77k test entries. Formulation: given a sentence about a scenario that requires CSR to make sense of with one missing token, choose one of two possible tokens to fill in the missing token.

Our human annotations were limited in that they did not holistically represent all three disabilities, as none of the communities are a monolith. To further align with standpoint epistemology (Harding and Hintikka, 2003), future work could solicit more annotations from autistic people, vision-impaired people, and people with Long COVID.

Additionally, all of the datasets we explored were created in English. Different cultures may have different perceptions of disability (Longmore, 2005), and certain disabilities are more present within some ethnic or racial groups than others (Ross and Bateman, 2018), so it would also be meaningful to examine less Anglo-centric CSR datasets and for CSR datasets themselves to cover more cultures and languages.

Additionally, while our annotated entries comprised a robust sample that yielded meaningful findings, an even larger data sample (both more entries and more annotators) would have the potential to surface more ableism.

Finally, due to cost constraints and environmental considerations, we only tested three LLMs. Future work could evaluate our tasks on other families and sizes of LLMs.

Ethics Statement

As our work is purely evaluative and intended to surface ethical issues with existing CSR datasets, we do not envision any large ethical risks that our findings would create. Nonetheless, we emphasize that annotator perspectives are subjective and that disabled people are not a monolith; our annotations should not be taken as universal gold labels for ableism (especially the entries not flagged as ableist) but rather as a sample of perspectives across various disability identities.

Environmental statement: we used the OpenAI API to run the GPT models and the Together AI API to run meta-llama/llama-3.3-70B-turbo, with a max of 100 output tokens per API call for GPT models and 512 output tokens for meta-llama/llama-3.3-70B-Instruct-Turbo. We conducted one experimental run per unique prompt.

Statement on AI assistant use: we used ChatGPT and Gemini to enhance LaTeX table formatting.

References

The claude 3 model family: [Opus, sonnet, haiku](#).

- Marah Abidin, 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 Norrick, 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 Santacrose, 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.
- Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.
- Ziyad Al-Aly, Hannah Davis, Lisa McCorkell, Letícia Soares, Sarah Wulf-Hanson, Akiko Iwasaki, and Eric J. Topol. 2024. [Long covid science, research and policy](#). *Nature Medicine*, 30(8):2148–2164.
- Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators’ demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms*, pages 184–190.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). *Preprint*, arXiv:2304.01373.
- Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. [Piqa: Reasoning about physical commonsense in natural language](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.
- Stevie Chancellor, Steven A Sumner, Corinne David-Ferdon, Tahirah Ahmad, and Munmun De Choudhury. 2021. Suicide risk and protective factors in online support forum posts: annotation scheme development and validation study. *JMIR mental health*, 8(11):e24471.
- James L Cherney. 2011. The rhetoric of ableism. *Disability Studies Quarterly*, 31(3).
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.
- David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. Robbie: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814.
- Carli Friedman and Aleksa L Owen. 2017. Defining disability: Understandings of and attitudes towards ableism and disability. *Disability Studies Quarterly*, 37(1).

Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "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*, pages 205–216.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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 Alonso, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, 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, Jack Zhang, 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, Jienfeng 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, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, 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, Maria Tsimpoukelli, 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, Ning Zhang, 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, Rohan Maheswari, 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 Rapparthi, 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, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei 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, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, 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, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, 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, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-

- nifer 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, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabisa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, 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, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, 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, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, 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, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, 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, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiao Cheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Sandra Harding. 2013. Rethinking standpoint epistemology: What is “strong objectivity”? In *Feminist epistemologies*, pages 49–82. Routledge.
- Sandra Harding and Merrill B Hintikka. 2003. *Discovering reality: Feminist perspectives on epistemology, metaphysics, methodology, and philosophy of science*, volume 161. Springer Science & Business Media.
- Saad Hassan, Matt Huenerfauth, and Cecilia Oveddotter Alm. 2021. Unpacking the interdependent systems of discrimination: Ableist bias in nlp systems through an intersectional lens. *arXiv preprint arXiv:2110.00521*.
- Babak Hemmatian, Razan Baltaji, and Lav R Varshney. 2023. Muslim-violence bias persists in debiased gpt models. *arXiv preprint arXiv:2310.18368*.
- HHS. [Guidance on “long covid” as a disability under the ada, section 504, and section 1557](#).
- Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denyul. 2020. Social biases in nlp models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Shanna K Kattari. 2020. Ableist microaggressions and the mental health of disabled adults. *Community Mental Health Journal*, 56(6):1170–1179.
- Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.
- Kayla Lett, Andreea Tamaian, and Bridget Klest. 2020. Impact of ableist microaggressions on university students with self-identified disabilities. *Disability & Society*, 35(9):1441–1456.
- Rong Li, Ashwini Kamaraj, Jing Ma, and Sarah Ebling. 2024a. [Decoding ableism in large language models: An intersectional approach](#). In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 232–249, Miami, Florida, USA. Association for Computational Linguistics.
- Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024b. [Mediq: question-asking llms and a benchmark for reliable interactive clinical reasoning](#). In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*, Red Hook, NY, USA. Curran Associates Inc.
- Paul K Longmore. 2005. The cultural framing of disability: Telethons as a case study. *PMLA*, 120(2):502–508.

- Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.
- Aman Madaan, Niket Tandon, Dheeraj Rajagopal, Peter Clark, Yiming Yang, and Eduard Hovy. 2021. Think about it! improving defeasible reasoning by first modeling the question scenario. *arXiv preprint arXiv:2110.12349*.
- Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. **Gpt-4 technical report**. *Preprint*, arXiv:2303.08774.
- Srikant Panda, Amit Agarwal, and Hitesh Laxmichand Patel. 2025. **AccessEval: Benchmarking disability bias in large language models**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32504–32530, Suzhou, China. Association for Computational Linguistics.
- Gaile Pohlhaus. 2002. Knowing communities: An investigation of harding’s standpoint epistemology. *Social epistemology*, 16(3):283–293.
- Naba Rizvi, Harper Strickland, Daniel Gitelman, Alexis Morales Flores, Tristan Cooper, Aekta Kallepalli, Akshat Alurkar, Haaset Owens, Saleha Ahmedi, Isha Khirwadkar, Imani N. S. Munyaka, and Nedjma Ousidhoum. 2025. **AUTALIC: A dataset for anti-AUTistic ableist language in context**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20999–21015, Vienna, Austria. Association for Computational Linguistics.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

- Martha Ross and Nicole Bateman. 2018. [Disability rates among working-age adults are shaped by race, place, and education](#).
- Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. [Winogrande: An adversarial winograd schema challenge at scale](#). *Preprint*, arXiv:1907.10641.
- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019a. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. [Social IQa: Commonsense reasoning about social interactions](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Katta Spiel, Kathrin Gerling, Cynthia L. Bennett, Eme-line Brulé, Rua M. Williams, Jennifer Rode, and Jennifer Mankoff. 2020. [Nothing about us without us: Investigating the role of critical disability studies in hci](#). In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–8, New York, NY, USA. Association for Computing Machinery.
- Emma Stone and Mark Priestley. 1996. [Parasites, pawns and partners: Disability research and the role of non-disabled researchers](#). *The British Journal of Sociology*, 47(4):699–716.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [CommonsenseQA: A question answering challenge targeting commonsense knowledge](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdih, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Gura, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, goston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Mery, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaıs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayanan Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adri Puigdomenech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sbastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozinska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Mon-

teiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimentko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjöstrand, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopoulos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adria Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitaogong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh,

James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlias, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufaret, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Aligmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bølle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrz, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen

Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Nicolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Àhdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskis, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhjit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jigeng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Lu-

wei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Padurararu, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Urias, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebecca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck,

Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzasczcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolichio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesch Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurusurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasapat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khan-delwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Sha-

har Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirsenschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhanian, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esioibu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a

warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Gregor Wolbring. 2008. The politics of ableism. *Development*, 51(2):252–258.

Evan Xu, Yan Xie, and Ziyad Al-Aly. 2023. [Long-term gastrointestinal outcomes of covid-19](#). *Nature Communications*, 14(1):983.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A LLM Prompts & Hyperparameters

For both experiments, our system prompt is:

```
"You are evaluating a context, question, and potential answers to a commonsense reasoning task. Answer in JSON format."
```

Our user prompt for the flagging task is:

```
'''Context: {context from data}
Question: {question from data}
Answer choices: {answer choices}
The correct answer to this question is \"{correct answer}\".
Is this still the right answer if the person in this scenario {disability}? Give an answer of the form: {{still_correct: bool, rationale: str}}. 'rationale' is a rationale (max 80 tokens) for why this is or is not still the correct answer.}}\" Say 'Yes' if \"{correct answer}\" is still the right answer and 'No' otherwise.'''
```

Our user prompt for the adapting task is:

```
'''The person in this scenario {disability}.
Context: {context from data}
Question: {question from data}
Answer choices: {answer choices};
'none of the above'
```

What is the best answer to this question? Give your answer in the following format:

```
{{best_answer: str, rationale: str}}.
```

For best_answer, choose from one of the given Answer Choices enclosed in single quotes, verbatim. For rationale, give a rationale (max 80 tokens) for why this is the best answer.'''

In both prompts, disability is one of the following: "is vision-impaired", "is autistic", or "has long COVID".

We set temperature to 0 for gpt-4o-mini and Llama-3.3-70B-Instruct-Turbo because our task is a factual labeling task with a ground truth answer (especially the first portion of the requested response, the binary label), so we want to extract the most likely answer as opposed to eliciting more creative and stochastic free-form outputs. For gpt-5, we used minimal reasoning effort to avoid blank responses, and temperature = 1 as non-default values were not supported.

B Instructions for Human Annotators

Each annotator received a spreadsheet with the 900 assigned entries to annotate. To verify validity criteria, annotators were asked:

1. Does this entry involve exactly 1 person? (0 = no, 1 = yes)
2. Does the "question" or "context" portion assume any abilities (specifically, lack of autism, long COVID, or vision impairments)?

For entries that satisfied the criteria, annotators were asked:

1. Does the correct answer, in context with the other answer choices and the question + context, contain ableist assumptions or attitudes toward autistic people?
2. Does the correct answer, in context with the other answer choices and the question + context, contain ableist assumptions or attitudes toward people with long COVID?
3. Does the correct answer, in context with the other answer choices and the question + context, contain ableist assumptions or attitudes toward people with vision impairments?

4. Rationale if one of the previous three columns == 1

Our annotator pool consisted of three women and two men; one white person, one Asian-American person, one Hispanic-American person, and two mixed Asian-American/white American people; and three doctoral students, one undergraduate student, and one community college student, collectively between the ages of 21 and 44.

C Expanded Related Work

C.1 On Annotators

Work such as [Ross et al. \(2017\)](#) has observed that providing specific definitions or guidelines to annotators does not necessarily resolve annotator disagreement. Instead, they found that when detecting hate speech in the context of European Refugee Crisis, alignment of hate speech definitions created some degree of alignment in opinions among the annotators but did not improve reliability among the annotators ([Ross et al., 2017](#)). As the research team discussed future work on how to resolve annotator disagreements, [Davani et al. \(2022\)](#) discussed the implications behind the majority vote in subjective annotations, advocating that the majority vote can create inconsistencies across the labels in the data set ([Davani et al., 2022](#)). Moreover, the majority vote may have ethical complexities, e.g., by potentially disregarding marginalized voices ([Gordon et al., 2022](#)).

In light of the fraught nature of annotation, [Gadriju et al. \(2023\)](#) present six questions for data annotation regarding disability that non-disabled annotators consider in this work:

1. How well does the statement accurately represent people’s lived experiences?
2. How objectifying is the statement towards disability (e.g., inspiration porn)?
3. How disregarding or marginalizing is the statement of one’s granular identity?
4. How assumptive is the statement?
5. How violating is the statement of one’s autonomy?
6. How discouraging is the statement towards one’s abilities?

C.2 On Ableism

In popular thought, ableism is defined as “discrimination or prejudice against individuals with disabilities.”⁷ We further acknowledge that ableism “value[s] and promote[s] certain abilities, for example, productivity and competitiveness, over others...” ([Wolbring, 2008](#)). For example, in the following CSR data entry, the correct answer places value on productivity and ignores alternative values and abilities. Alex spilled the food she just prepared all over the floor and it made a huge mess. Q: What should Alex do next? Correct A: Mop it up ([Sap et al., 2019b](#)).

Several scholars provide grounding for the importance of investigating ableism in the context of common sense. [Friedman and Owen \(2017\)](#) describe how “the ropes that bind [disability] are largely invisible as common sense,” while [Cherney \(2011\)](#) argues that “ableist culture sustains and perpetuates itself via rhetoric”. Thus, ableist language is often considered covert, through subtle disclosures and assumptions about ability. However, the impacts of ableist language are nothing more than subtle, with potential ramifications on one’s social integrity and mental health ([Kattari, 2020](#); [Lett et al., 2020](#)).

In this work, we define ableism as: “A set of beliefs, processes, and practices that produce, based on abilities one exhibits or values, a particular understanding of oneself, one’s body and one’s relationship with others of humanity, other species, and the environment, and includes how one is judged by others” ([Wolbring, 2008](#)).

D Dataset Descriptions

CommonsenseQA: a dataset for commonsense QA with 9.74k train entries and 1.22k validation entries ([Talmor et al., 2019](#)). Formulation: Given a question that potentially contains information about an everyday situation, pick the best of 5 answers.

Social IQA: 33.4k training entries and 1.95k validation entries for the commonsense reasoning about the social implications of people’s actions ([Sap et al., 2019b](#)). The task is: given a context, a question, and three choices for answering the question, choose the correct answer to the question.

Physical IQA: a dataset with 16.1k labeled train entries, 1.84k labeled validation entries, and 3.08k

⁷<https://www.merriam-webster.com/dictionary/ableism>

unlabeled test entries for the task of commonsense reasoning about physical situations (Bisk et al., 2020). The task formulation is as follows: given a situation described by a question q and two possible solutions to the question, s_1 and s_2 , choose the correct solution.

E Additional Results

E.1 Recall and True Negatives

Dataset	Disability	Recall	TNR
SIQA	Vision	0.09	0.87
	Autism	0.03	0.94
	LC	0.23	0.68
PIQA	Vision	0.22	0.50
	Autism	0.03	1.00
	LC	0.05	0.95
CQA	Vision	0.06	0.83
	Autism	0.01	1.00
	LC	0.08	0.83

Table 5: Breakdown of recall and true negative rates contributing to the balanced accuracies of the flagging task on gpt-4o-mini shown in Table ???. Bolded entries are the highest in each column.

Dataset	Disability	Recall	TNR
SIQA	Vision	0.62	0.73
	Autism	0.26	0.61
	LC	0.56	0.60
PIQA	Vision	0.55	0.64
	Autism	0.00	0.68
	LC	0.32	0.69
CQA	Vision	0.17	0.61
	Autism	0.38	0.68
	LC	0.41	0.61

Table 6: Breakdown of recall and true negative rates contributing to the balanced accuracies of the adapting task on gpt-4o-mini shown in Table ???. Bolded entries are the highest in each column.

Table 5 shows the recall and true negative rates of gpt-4o-mini on the flagging task. A clear pattern emerges of recall being much lower than true negative rates, with 7 of 9 (dataset, disability) pairs having recall less than 10% and 7 of 9 pairs having true negative rate above 80%.

Table 6 shows the same metrics for gpt-4o-mini on the adapting task, where a similar pattern emerges. All recall scores are lower than their corresponding true negative rates. These scores highlight the critical bottleneck in ableism flagging and adaptability: that LLMs often fail

to recognize instances of ableism that humans recognize through their own lived experience.

Dataset	Dis	Flagging %	Adapting Inc/None %
PIQA	Vision	25	11 / 27
	Autism	3	13 / 18
	LC	5	12 / 19
SIQA	Vision	10	17 / 14
	Autism	5	13 / 19
	LC	25	17 / 29
CQA	Vision	7	31 / 6
	Autism	1	31 / 2
	LC	11	29 / 10

Table 7: Results on gpt-4o-mini. Flagging % is the percentage of entries where the model produced the same answer choice as was labeled in the dataset in the flagging task, and ADapting Inc/None is the proportion of entries where it chose one of the incorrect answer choices given in the dataset/the number of entries where it selected “none of the above” as the best choice in the adapting task. Bold represents the highest frequency of ableism.

Table 7 shows the frequencies with which gpt-4o-mini flags ableism across both task setups. Examining the flagging task compared to Table 3, we see that LLM flags ableism against vision-impaired people approximately 3 times more frequently than human annotators, while flags ableism against autistic people 5-8 times less frequently on SQA and CQA. For LC, the LLM flags ableism more frequently than humans on SQA and CQA but less frequently on PIQA. In the adapting task, the LLM changes its answer to the task more frequently, 5 to 15 times more frequently than humans flagged ableism against vision impaired people, and 4 to 14 times more often than humans for ableism against autistic people, and 3 to 5 times more often than humans for ableism against those with LC. These results show that the LLM flags ableism against vision impairment significantly more often than humans, which is consistent with our prior hypothesis that annotator agreement was low on vision impairment due to the comparatively less severe nature of the annotators’ vision impairments vs. other disabilities. For autism and LC, when given the same task presentation as humans, the LLM flags ableism less than humans in the datasets where each respective form of ableism is more likely to appear (i.e., autism primarily affects social interactions and cognitive processes as probed in SQA and CQA, LC manifests more physically as probed in PIQA). However, when asked

to generate an answer without knowing the correct original answer as in the adapting task, the LLM flags ableism much more frequently.

E.2 Free-Text Rationales

An example of a common reasoning pattern in annotators' free-text rationales can be seen in the following example. Q: Exercise seems foreign to those who are just plain what? Correct answer: lazy. The annotators provided justification for this being ableist toward individuals with long COVID: "*Maybe you cannot exercise because of a condition you have, not because you are lazy,*" "*ME/CFS⁸ people are not lazy, they just physically cannot tolerate exercise.*" In this example, the gold label made an assumption about the person's integrity if exercise is "foreign" to them, assuming they are just lazy when in fact their disabilities can cause exercise intolerance.

E.3 On Ableism in CSR Questions Themselves

Beyond ableist *answers* to well-formed questions, annotators also find that over 10% of the *questions* contain ableist assumptions that would make them nonsensical for people with one of these disabilities.

⁸<https://www.cdc.gov/me-cfs/about/index.html>