Do LLMs Authentically Represent Affective Experiences of People with Disabilities on Social Media?

Marco Bombieri[1,*], Simone Paolo Ponzetto[2] and Marco Rospocher[1]

[1]University of Verona, Lungadige Porta Vittoria, 41, 37129 Verona, Italy
[2]University of Mannheim, B6, 26, D-68159 Mannheim, Germany

Abstract

This paper investigates how Large Language Models (LLMs) represent the affective experiences of individuals with disabilities on social media. We simulate posts using LLMs and compare them to authentic user-generated content in English, collected from disability-related subreddits, focusing on sentiment, emotion, and indicators of depression. Our analysis reveals that LLMs tend to produce overly positive and idealized portrayals, often failing to capture the complexity and nuance of disabled individuals' emotional expressions. These misrepresentations underscore broader concerns about the limitations of LLMs in authentically reflecting the lived experiences of marginalized communities.

Keywords

Large Language Models, Representation, Disability, Bias

1. Introduction

Recent studies have shown that computational models of language, trained on real-world data, reflect and amplify harmful societal biases, often disproportionately affecting marginalized communities [1, 2, *inter alia*]. This can lead to psychological harm, unhappiness, and, in some cases, suicide attempts [3]. The increasing use of Large Language Models (LLMs) has exacerbated the risks related to this issue, potentially spreading these representational harms further [4]. In response, researchers have proposed methods to mitigate these biases. For example, recent LLMs have incorporated de-biasing techniques and AI guards (e.g., Inan et al. [5]) that block offensive questions and adjust responses to be non-toxic and positive. However, recent work on studying the depiction of personas from marginalized groups of LLMs indicates that many biases are concealed even in texts containing words with a positive sentiment, which can still offend their sensitivities and lead to pernicious positive portrayals [6]. Moreover, in the specific case of disability, excessive positivity can be counterproductive to inclusion: some members of the disability community express dissatisfaction when they are portrayed in an excessively and pathetically positive and optimistic manner: according to them, this form of optimism reinforces what is known as "inspiration porn" [3, 7, 8] which has the negative consequence of dehumanizing individuals with disabilities, leading society to praise their efforts rather than working toward tangible solutions that alleviate the often strenuous challenges they face in survival through accessible political and social policies.

In this paper, we thus examine how current LLMs portray individuals with disabilities[1] from an affective perspective. Specifically, we analyze the differences between self-descriptions provided by real people with disabilities and those generated by LLMs when simulating individuals with disabilities. Our focus is on assessing the sentiment, emotional tone, and levels of depression in these descriptions, with the aim of understanding how authentically LLMs represent the emotional experiences of people with disabilities and identify differences and patterns in the affective portrayal of disability in AI-generated content.

Our work aims to deepen discussions on how LLMs should authentically represent disability, a topic that has received comparatively less attention in NLP literature [3], despite the frequent discrimination faced by disabled individuals [9, 10]. Specifically, we address the following Research Question (RQ):

> Can LLMs authentically represent the affective experiences of people with disabilities on social media?

Answering the above RQ, we offer the following contributions:

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
marco.bombieri_01@univr.it (M. Bombieri); ponzetto@uni-mannheim.de (S. P. Ponzetto); marco.rospocher@univr.it (M. Rospocher)
0000-0002-8607-8495 (M. Bombieri); 0000-0001-7484-2049 (S. P. Ponzetto); 0000-0001-9391-3201 (M. Rospocher)

[1]In this paper, we primarily use people-first language (e.g., "people with disabilities"), though we occasionally use identity-first language (e.g., "disabled people", "non-disabled people") based on sentence structure. We recognize that preferences for people-first or identity-first language vary among individuals. we intend not to offend or diminish anyone's perspective.

C1. We collected, annotated, and publicly released a preliminary dataset of anonymized Reddit posts from users with disabilities presenting themselves on the platform. Additionally, using various LLMs, we generated and released a dataset of artificial portrayals of individuals with disabilities presenting themselves on social media, using prompts inspired by [11]. Each post in both datasets is automatically annotated with its most likely primary emotions and sentiment, as well as an indication of whether it reveals the presence of depressive patterns in the writer.

C2. We compared web-collected posts with those generated by LLMs to study how models represent individuals with disabilities from an affective point of view, identifying differences between real-world and AI-generated portrayals.

Our findings emphasize the need to expand research on stereotypes to address both negative ones and positive idealizations, as both can harm marginalized groups. Furthermore, the analysis of the dataset on people with disabilities reveals significant challenges they frequently face, often associated with negative emotions or depressive symptoms, a fact already observed in literature [12]. Experiments also show that LLMs tend to minimize these aspects when portraying people with disability and substitute them with a more socially desirable narrative.[2]

2. Related Work

LLMs and Fairness. Recent advancements in LLMs have transformed text processing and generation, increasingly shaping social interactions. However, these models can perpetuate harmful stereotypes and biases [4], inheriting issues from uncurated internet data, such as misrepresentations, derogatory language, and biased associations [13, 6, 14, 1, 2]. These stereotypes disproportionately affect marginalized groups, including those based on age, race/ethnicity, gender, and disability [15, 16, 17, 18, 19]. As awareness of these misrepresentations grows, research has focused on bias and stereotypes evaluation, mitigation methods, and datasets to address them [4]. However, despite 1.3 billion people living with disabilities [20], there is limited research on stereotypes regarding disability representation in LLMs [21, 22]. Furthermore, existing datasets like BBQ [23], HolisticBias [19], and PANDA [24] address disability representation partially, lacking a comprehensive range of impairments and analysis.

[2]The code and the dataset are available at: https://github.com/marcobombieri/LLM-disability-representation

Bias against people with disabilities. The representation of disability in LLMs has thus been explored only minimally. Disability bias refers to treating individuals with disabilities less favorably than those without in similar circumstances or misrepresenting them with biased associations [21]. Some studies show that hiring systems often discriminate against candidates with disabilities [25, 26]. In particular, Glazko et al. [26] highlights that even GPT-4 shows bias in suggesting job candidates. Venkit et al. [21] and Hutchinson et al. [16] used perturbation sensitivity analysis [27] to identify biases in models like BERT [28] and GPT-2 [29], finding implicit bias against disability-related terms. [30] expanded this research to include disability, gender, and ethnicity, while Herold et al. [31] found BERT frames disabilities mainly in medical terms. Recent work by Li et al. [32] suggests newer models like GPT-3.5 and GPT-4 offer less biased portrayals of disabilities.

LLM-based portrayals and human simulation. A related research trend is human simulation, where LLMs are assessed on their ability to replicate human behavior, a concept introduced by the Turing Experiment [33]. This is applied to simulate behavior in various social and political settings [34, 35] and to identify stereotypes [11, 6]. Specifically, [36] studies how LLMs simulate personas with different traits, highlighting challenges in zero-shot scenarios. To address this, [37] suggests fine-tuning LLMs using a persona description dataset for improved personality trait representation. Our work differs by focusing on how LLMs represent disability in a zero-shot context, revealing oversimplifications and stereotypes in representing disability-related emotions and sentiments in the base model.

3. Methods

3.1. A dataset of LLMs-generated portrayas of people with disabilities

Using a set of prompts inspired by Kambhatla et al. [11] and Cheng et al. [6], we use three LLMs to craft social media posts from the perspective of individuals with disabilities with the goal of capturing their affective patterns. In particular, we employ Mixtral-8B (Mixtral-8x7B Instruct, quantized to 4 bits using GPTQ [38] due to hardware limitations), GPT-4o-mini, and Gemini-1.5F (the Flash variant). For the latter two models, we utilized their paid APIs. The models are used with the temperature set to 1.0 to guarantee the variability and randomness of the responses generated.

Each prompt asks the models to introduce themselves as an individual with a disability and share experiences related to disability or general life experiences such as ex-

[PERSON]	*[ACTIVITY]*
P1: *a person with a disability*	**A1**: *sharing experiences related to your disability*
P2: *a person with autism*	**A2**: *sharing the emotions you felt today*
P3: *a person with blindness*	**A3**: *sharing the thoughts you had today*
P4: *a person with deafness*	**A4**: *sharing the activities you did today*
P5: *a person with cerebral palsy*	**A5**: *asking the community a question or suggestion*
P6: *a person with depression*	**A6**: *commenting on today's events*

Table 1
Possible values for *[PERSON]* and *[ACTIVITY]* in our prompt template: *"Imagine you are [PERSON]. Write a social media post introducing yourself and [ACTIVITY]."*

pressions of emotions, feelings, or thoughts, descriptions of daily activities, questions for the community, requests for suggestions, or commentary on current events, i.e., the typical activities a user can do on a social media platform [39]. We opted to keep the prompts as general as possible following the motivations discussed in [6], since more detailed prompts may direct the model toward a specific topic and introduce further stereotypes. In more detail, all the prompts follow the template:

> *"Imagine you are [PERSON]. Write a post on social media introducing yourself and [ACTIVITY]."*

where *[PERSON]* and *[ACTIVITY]* can be one of those defined in Table 1.

The combination of P1-P6 with A1-A6 aims to generate posts from the perspective of individuals with different types of disabilities or impairments. Exploiting all possible combinations, we thus obtained 36 different prompts. Each prompt is submitted 10 times to take into account the output variability of the models, thus obtaining, for each LLM, a collection of 360 posts of artificial portrayals of people with disabilities. We call LLMD_{GPT}, LLMD_{GEM}, and LLMD_{MIX} the datasets containing the posts generated by GPT-4o-mini, Gemini-1.5F, and Mixtral-8B, respectively. In this preliminary work, we narrow our focus to the disabilities examined in similar studies, such as [26], resulting in six alternative options (P1–P6) for *[PERSON]*.

3.2. A dataset of people with disabilities' self-descriptions

In addition to the datasets described in Section 3.1, we collected posts from six disability-related subreddits. We began with the general subreddit `r/disability`[3], which offers diverse discussions on disability-related topics and ranks among the top 2% by size. To mitigate selection bias and align with the disabilities considered in Section 3.1, we added five focused subreddits: `r/blind`[4], `r/autism`[5], `r/depression`[6], `r/deaf`[7], and `r/celebralpalsy`[8]. These subreddits aim to foster community and exchange among disabled individuals. We included posts published until 2024 containing textual content, excluding empty posts or those with only links, images, or videos. Using Mixtral-8B and the below prompt, we filtered for first-person posts from users self-identifying as disabled, excluding content from caregivers, professionals, or others:

> *You are a text classifier operating on social media posts. You must classify posts into two disjoint classes, "1" or "2". Your answer must be in the format: "predictedClass;explanation," where "predictedClass" can be "1" or "2," and "explanation" briefly describes why you have chosen that class. Separate "predictedClass" from "explanation" with the string ";". Do not add other text. A post belongs to class "1" if: (the author of the post writes about himself/herself in the first person) AND (the author of the post explicitly mentions his/her own disability/illness). A post belongs to class "2" otherwise. Follow the post you have to analyze:*
> {word}

From the filtered results, we randomly sampled 450 posts from `r/disability` and 220 from each of the disability-specific subreddits. Three annotators then manually reviewed all these posts, removing those wrongly annotated as relevant by the LLM. The final dataset, REDd, includes 352 posts from `r/disability`, 165 from `r/blind`, 174 from `r/autism`, 204 from `r/depression`, 171 from `r/deaf`, and 183 from `r/cerebralpalsy`.[9] To ensure annotation quality, 50

[3]Subreddit `r/disability`: https://www.reddit.com/r/disability/ [Last access: 2025-05-16]

[4]Subreddit `r/blind`: https://www.reddit.com/r/blind/ [Last access: 2025-05-16]

[5]Subreddit `r/autism`: https://www.reddit.com/r/autism/

[6]Subreddit `r/depression`: https://www.reddit.com/r/depression/

[7]Subreddit `r/deaf`: https://www.reddit.com/r/deaf/

[8]Subreddit `r/cerebralpalsy`: https://www.reddit.com/r/cerebralpalsy/

[9]Our goal is not to develop an LLM for post classification, but to

Dataset	Description	# Post	Avg. Tokens
LLMD_{GEM}	Dataset of posts generated by **GEMINI-1.5F** when representing a person with a disability.	360	243.01
LLMD_{GPT}	Dataset of posts generated by **GPT-4O-MINI** when representing a person with a disability.	360	221.66
LLMD_{MIX}	Dataset of posts generated by **MIXTRAL-8B** when representing a person with a disability.	360	247.97
REDD	Dataset of posts of Reddit users with disabilities.	1,250	207.55
LLMD	Dataset created by concatenating LLMD_{GEM}, LLMD_{GPT}, and LLMD_{MIX}.	1,080	237.55

Table 2
Summarization of datasets collected in this paper, together with the number of posts they contain and the average number of tokens per post.

posts were independently labeled by three annotators, achieving a Fleiss' Kappa of 0.875, indicating very high agreement [40]. Table 2 summarizes the obtained datasets and their sizes that are in line with state-of-the-art studies [6].

3.3. Comparison metrics

To address our research question, we aim to perform a pairwise comparison of the previously described datasets, i.e., the LLM-generated portraits (Section 3.1) and human descriptions from Reddit users (Section 3.2) using metrics descriptive of the affects of an individual. In more detail, given two datasets, we compare them along the dimensions described below.

Sentiment. The predominant *sentiment* of each post p is computed using VADER [41], which assigns a sentiment score $S(p) \in [-1, +1]$. Following VADER indications, a post is classified as positive if $S(p) > 0.05$, negative if $S(p) < -0.05$, and neutral otherwise. For a dataset $P = [p_1, \ldots, p_N]$ of N posts, we compute the number of positive, negative, and neutral posts:

$$N_{\text{positive}} = |\{p_i \mid S(p_i) > 0.05\}|,$$
$$N_{\text{negative}} = |\{p_i \mid S(p_i) < -0.05\}|,$$
$$N_{\text{neutral}} = |\{p_i \mid -0.05 <= S(p_i) <= 0.05\}|.$$

We then compute the relative frequency of sentiment-loaded posts:[10]

$$P_{\text{positive}} = \frac{N_{\text{positive}}}{N},\ P_{\text{negative}} = \frac{N_{\text{negative}}}{N}.$$

Emotions. The distribution of *emotions* emerging from a dataset using the NRC Word-Emotion Association Lexicon (EmoLex) [42], namely *anger*, *fear*, *anticipation*, *trust*, *surprise*, *sadness*, *joy* and *disgust*. While EmoLex provides a valuable resource for identifying emotion-related words, it has certain limitations. Specifically, it is based solely on word-level counts from the lexicon. It does not account for contextual factors such as negations, word dependencies, or the broader semantic structure of the text. Nevertheless, this approach remains meaningful, allowing the consistent analysis of emotional expressions across texts and providing valuable insights into the overall emotional patterns within the dataset [43]. Let $P = \{p_1, p_2, \ldots, p_N\}$ represent the dataset with its set of N posts. For each post p_i, we calculate the number of words associated with each emotion $e \in E$, denoted by w_{e,p_i}, where w_{e,p_i} is the number of words in post p_i that are associated with emotion e. If a word is linked to multiple emotions, all associated emotions are considered. The proportion ρ_{e,p_i} of words in post p_i associated with emotion e is given by:

$$\rho_{e,p_i} = \frac{w_{e,p_i}}{w_{p_i}}$$

where w_{p_i} is the total number of words in post p_i that are linked to any emotion. At the dataset level, the average proportion of each emotion across all posts is computed as:

$$\bar{\rho}_e = \frac{1}{N} \sum_{i=1}^{N} \rho_{e,p_i}.$$

Depression. The indication of the presence of *depression* as determined by the best-performing model from the Shared Task on *Detecting Signs of Depression from Social Media Text* at LT-EDI-ACL2022 [44].

Let $p_{i,l}$ denote the predicted depression label for a given post p_i, where:

$$p_{i,l} \in \left\{ \begin{array}{l} l_1 = \textit{no depression}, \\ l_2 = \textit{moderate depression}, \\ l_3 = \textit{severe depression} \end{array} \right\}.$$

To analyze the distribution of labels across the dataset,

compile a dataset of posts by people with disabilities to support our analysis; the LLM (78% accuracy) was used solely to assist filtering.

[10] Posts with scores between -0.05 and 0.05 are considered neutral. Since REDD is the only dataset containing neutral posts — and only two such posts — we chose to focus the following analysis exclusively on positive and negative posts.

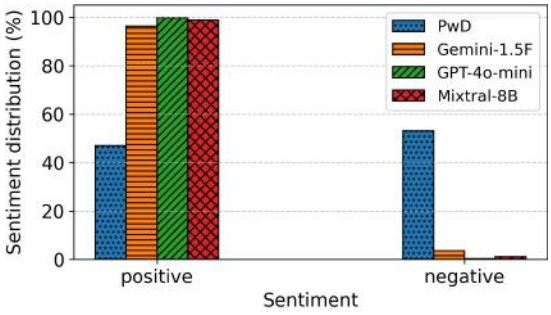

Figure 1: Comparison of sentiment between posts from people with disabilities (PwD) on Reddit (REDD Dataset) and posts generated by GEMINI-1.5F (LLMD$_{GEM}$ Dataset), GPT-4O-MINI (LLMD$_{GPT}$ Dataset), and MIXTRAL-8B (LLMD$_{MIX}$ Dataset).

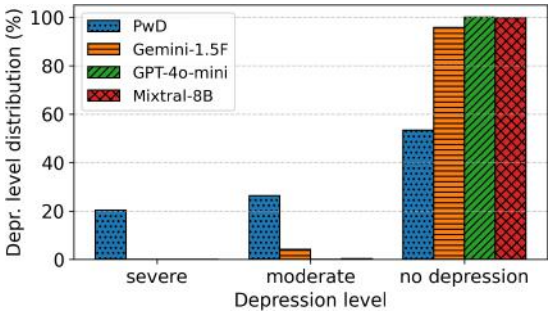

Figure 2: Comparison of depression levels between posts from people with disabilities (PwD) on Reddit (REDD Dataset) and posts generated by GEMINI-1.5F (LLMD$_{GEM}$ Dataset), GPT-4O-MINI (LLMD$_{GPT}$ Dataset), and MIXTRAL-8B (LLMD$_{MIX}$ Dataset).

we define the proportion of each label $l \in \{l_1, l_2, l_3\}$ as:

$$P(l) = \frac{N_l}{N},$$

where N represents the total number of posts in the dataset, and N_l is the number of posts $p_{i,l}$ assigned to label l.

Sentiment, emotion, and depression analyses offer quantitative insights into emotional tone and mental health indicators. These analyses enable affective comparisons with LLM-generated texts and provide a preliminary valuable clues about how LLMs represent individuals with disabilities.

In our setting, to address our RQ, we perform sentiment, emotion, and depression analysis on LLMD$_{GEM}$, LLMD$_{GPT}$, LLMD$_{MIX}$, and REDD, comparing the first three datasets generated by the LLMs with the data from people with disabilities (REDD).

4. Results and Discussions

Figures 1, 2, and 3 illustrate the differences between the posts in REDD and those LLM-generated, i.e., those collected in LLMD$_{GPT}$, LLMD$_{MIX}$, and LLMD$_{GEM}$, in terms of sentiment, depression level, and emotion, respectively.

Figure 1 shows that the three LLMs overwhelmingly generate posts with positive sentiment, ranging from 99.72% for GPT-4O-MINI (LLMD$_{GPT}$ dataset) to 96.39% for GEMINI-1.5F (LLMD$_{GEM}$ dataset). In contrast, actual Reddit posts (REDD dataset) present a starkly different picture, with 53.06% of posts exhibiting negative sentiment. This discrepancy suggests that LLMs systematically underrepresent the negative emotional tone often present in real discussions about disability. The tendency to default to positivity may create an artificial and potentially misleading portrayal of lived experiences.

Figure 2 further reinforces this pattern, as GPT-4O-MINI exhibits no signs of depression, and MIXTRAL-8B has only one post classified as "moderate depression" in the LLMD$_{MIX}$ dataset. GEMINI-1.5F shows slightly higher rates, with 4.17% of posts categorized as "moderate depression" and 95.83% as "not depression". Notably, the few instances of moderate depression detected in LLM-generated content occur only when the models explicitly attempt to portray individuals with depression—and even then, at very the very low rates indicated above. These results contrast sharply with the Reddit dataset, where 20.42% of posts are labeled as "severe depression" and 26.26% as "moderate depression". In the collected dataset, posts exhibiting symptoms of depression are present across all the subreddits. The substantial under-representation of depressive expressions in LLM-generated content suggests that these models fail to capture the full emotional depth of real-life disability narratives. By filtering out or minimizing negative expressions, LLMs risk misrepresenting the struggles and challenges discussed in real-world communities, substituting them with a more palatable narrative that aligns with a non-disabled, socially desirable perspective.

Figure 3 further highlights these discrepancies, showing that Reddit posts contain significantly more negative emotions, such as anger, disgust, fear, and sadness, while LLM-generated posts emphasize positive emotions, including joy, trust, surprise, and anticipation. This over-representation of positivity suggests that LLMs adopt an overly optimistic and sanitized perspective on disability, potentially reinforcing harmful biases related to inspiration porn. The lack of emotional diversity in LLM-generated content may contribute to an inaccurate or even dismissive portrayal of the emotional realities experienced by people with disabilities.

Overall, these preliminary findings suggest that LLMs fail to authentically replicate the emotional tone of real experiences of social media disabled users. Instead, they

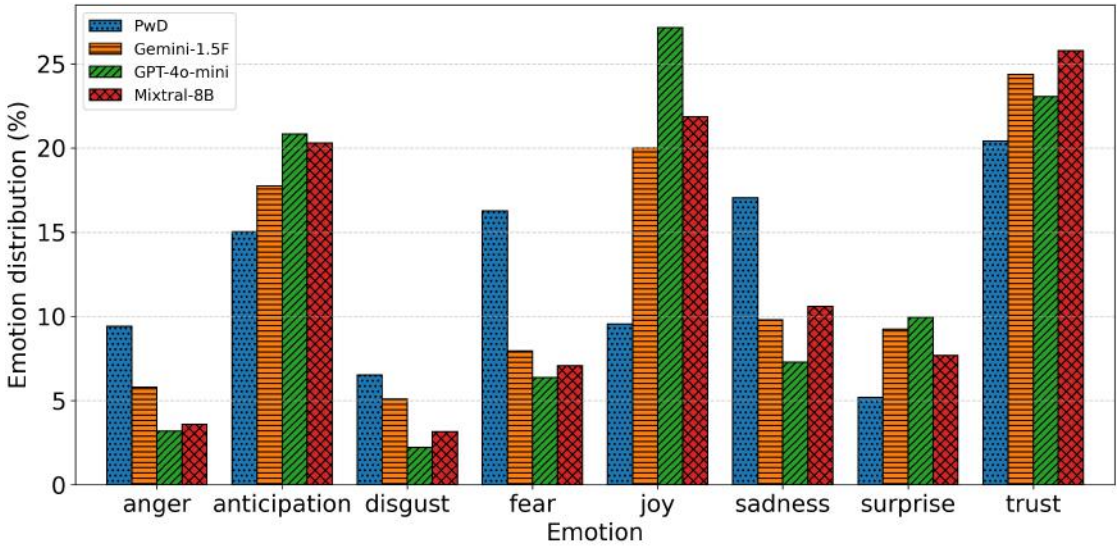

Figure 3: Comparison of emotions between posts from people with disabilities (PwD) on Reddit (REDD Dataset) and posts generated by GEMINI-1.5F ($LLMD_{GEM}$ Dataset), GPT-4O-MINI ($LLMD_{GPT}$ Dataset) and MIXTRAL-8B ($LLMD_{MIX}$ Dataset).

appear to spread a positivity bias, which may impact how disability is represented in AI-generated discourse.

To complement our quantitative metrics, we conduct a preliminary qualitative analysis of both LLM-generated and real posts, examining their structure and recurring themes. LLMs tend to frame disability through consistently positive lenses, emphasizing inclusion, accessibility, and triumph over adversity, with frequent use of words like *advocacy*, *inclusion*, *grateful*, *excited*, and *proud*. Follow an excerpt of a post generated by GPT-4O-MINI when representing a blind person:

> *I'm a **proud** member of the blind community. [...] One of my biggest passions is sharing my experiences and **advocating** for **accessibility** and **inclusion**. [...] I also want to highlight the **amazing** community I've found among fellow visually impaired individuals. We share stories, support one another, and **inspire** each other every day [...].*

In contrast, real posts by people with disabilities more often reference health, educational or financial struggles, using terms such as *pain*, *unemployed*, *bad*, and *anxiety*, **worse**, reflecting a broader emotional range and lived complexity.

Follow an excerpt of a post from `r/blind`:

> *I was born blind. Always been this way. From the time I was in high school, I began to have really **bad insecurities** about my blindness. [...] Growing up, **I hated** every blind person I went to school with. []. By the time I got to high school, **it just got worse and worse**. [...]*

In future research, we will expand this preliminary analysis with an in-depth qualitative and qualitative thematic analysis of posts.

Answer to RQ. The results reveal that the LLMs' affective descriptions of disability significantly differ from those expressed by real people with disabilities. LLM-generated texts largely emphasize positive sentiments and emotions, minimizing or entirely omitting the negative feelings that individuals with disabilities often experience. This tendency risks fostering a form of toxic positivity that overlooks the complex emotional landscape of disability, as highlighted by [45]. The analysis of REDD's posts, however, paints a starkly dangerous picture, where individuals with disabilities frequently express negative emotions such as anger, sadness, and fear. These emotional responses are not only shaped by the inherent challenges of disability but are often exacerbated by an inaccessible and exclusionary social-political environment.

5. Conclusions

In this paper, we investigated how LLMs represent disability from an affective point of view by comparing AI-generated portrayals with social media posts authored by individuals with disabilities. By leveraging a dataset

of Reddit posts and artificial portrayals generated by LLMs, we analyzed the emotional tone, sentiment, and depressive patterns of these texts. Our work contributes not only to a publicly available dataset but also to insights into the fundamental differences in how LLMs and real individuals describe disability, highlighting significant oversimplifications. Most specifically, through our experiments, we found that LLMs frequently idealize disability-related affective experiences, producing overly optimistic portrayals that ignore the complex realities and challenges faced by individuals with disability. In stark contrast, posts written by real individuals often convey more nuanced emotions, including negative feelings stemming from the intersection of their disabilities with inaccessible and non-inclusive societal systems.

This disconnect highlights the risk of toxic positivity, where overly optimistic portrayals diminish the real challenges faced by disabled individuals. Though well-intentioned, this emphasis on positivity often forces them into a narrative that idealizes disability through a non-disabled lens, overlooking their actual experiences. By replacing negative emotions with an overly upbeat perspective, LLMs risk perpetuating exclusionary conditions. Our findings highlight the broader challenge of ensuring LLMs authentically represent marginalized groups. While addressing negative stereotypes in AI is crucial, our study calls for a more nuanced approach that reflects the diverse realities of marginalized groups without reductive idealizations. This paper raises a critical question: should LLMs represent affective experiences in an exclusively optimistic, "good vibes only" manner, or should they strive for more authentic, emotionally complex portrayals that better reflect real human experiences?

In future work, we plan to test additional prompts and simulate a broader range of social media scenarios. We also plan to expand the collection of posts by including a wider range of subreddits, social media platforms, and languages. This will help capture a more diverse set of experiences from individuals with disabilities. We also aim to include a broader spectrum of disabilities and analyze how their representation varies across different categories. Additionally, we will enhance this study with thematic analysis methods to examine discourses related to disabilities in real and LLM-generated posts, identifying keywords that distinguish the two corpora—those written by disabled individuals and those generated by LLMs. A qualitative analysis will further complement this approach. Finally, comparing how LLMs portray individuals with disabilities versus the general population, following the methodology in [6], will offer deeper insights into these dynamics and help address the risk of oversimplification or misrepresentation.

Limitations

This paper is a preliminary work and thus has some limitations. First, we focused on a subset of disabilities to simplify the analysis. While this does not fully capture the complexity of the subject, it aligns with the approach taken in similar studies [26]. Second, we use lexicon-based tools to estimate emotions and sentiments, which may not always capture contextual nuances, potentially affecting the accuracy of the analysis. This methodology is, however, also employed in authoritative studies to ensure the method remains explainable and reproducible [6]. Furthermore, although we assume individuals who mention being disabled are indeed disabled, some may be bots or people pretending to be disabled. Finally, these findings are specific to the versions of the models and the dates on which they were tested (especially those accessed via API). As LLMs are updated and their guardrails evolve, these results may change.

Ethical and societal implications

This paper has a positive impact by shedding light on how disability is represented in zero-shot LLMs, emphasizing crucial ethical considerations. Current debiasing and representation models focus on “category” rather than “individual,” leading to potentially generalized, insensitive, or inappropriate responses. A model aiming to be inclusive must understand the personal experience of the individual represented. These models often fail to capture pain, suffering, and depression, substituting them with overly positive language. While optimism may be suitable in some cases, neglecting suffering flattens a key human experience. A “only good vibes” approach risks marginalizing those experiencing hardships, not just people with disabilities but anyone going through difficult times, exposing to the risk of inspiration porn. Therefore, these models must reflect the complexity of human emotions authentically and respectfully to foster genuine understanding, inclusion, and support. While addressing such personal topics may unintentionally cause misunderstandings, our intention is to promote constructive dialogue between technologists and humanists for more inclusive AI systems.

Data Availability

The code and the dataset are available at:
https://github.com/marcobombieri/LLM-disability-representation

Acknowledgments

This research has received funding from the University of Mannheim’s “Gastwissenschaftler*innenprogramm

Nachhaltigkeit", and the MUR funded 2023-2027 Project of Excellence "Inclusive Humanities: Perspectives for Development in the Research and Teaching of Foreign Languages and Literatures" of the Department of Foreign Languages and Literatures of the University of Verona. Part of this work was carried out within the Digital Arena for Inclusive Humanities (DAIH) Research Centre at the University of Verona. The authors gratefully acknowledge this support.

References

[1] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 4349–4357.

[2] T. Manzini, L. Yao Chong, A. W. Black, Y. Tsvetkov, Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019. doi:10.18653/v1/N19-1062.

[3] V. Gadiraju, S. K. Kane, S. Dev, A. S. Taylor, D. Wang, E. Denton, R. Brewer, "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 2023, Chicago, IL, USA, June 12-15, 2023, ACM, 2023, pp. 205–216. doi:10.1145/3593013.3593989.

[4] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, Computational Linguistics 50 (2024) 1097–1179. URL: https://doi.org/10.1162/coli_a_00524. doi:10.1162/coli_a_00524.

[5] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, M. Khabsa, Llama guard: Llm-based input-output safeguard for human-ai conversations, CoRR abs/2312.06674 (2023). URL: https://doi.org/10.48550/arXiv.2312.06674. doi:10.48550/ARXIV.2312.06674.

[6] M. Cheng, E. Durmus, D. Jurafsky, Marked personas: Using natural language prompts to measure stereotypes in language models, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, Association for Computational Linguistics, 2023, pp. 1504–1532. URL: https://doi.org/10.18653/v1/2023.acl-long.84. doi:10.18653/V1/2023.ACL-LONG.84.

[7] K. B. Ayers, K. A. Reed, Chapter 10 Inspiration Porn and Desperation Porn: Disrupting the Objectification of Disability in Media, Brill, Leiden, The Netherlands, 2022, pp. 90 – 101. doi:10.1163/9789004512702_014.

[8] J. Grue, The problem with inspiration porn: a tentative definition and a provisional critique, Disability & Society 31 (2016) 838–849. doi:10.1080/09687599.2016.1205473.

[9] L. VanPuymbrouck, C. Friedman, H. A. Feldner, Explicit and implicit disability attitudes of healthcare providers., Rehabilitation psychology (2020).

[10] G. Szumski, J. Smogorzewska, P. Grygiel, Attitudes of students toward people with disabilities, moral identity and inclusive education—a two-level analysis, Research in Developmental Disabilities 102 (2020) 103685. URL: https://www.sciencedirect.com/science/article/pii/S0891422220301153. doi:https://doi.org/10.1016/j.ridd.2020.103685.

[11] G. Kambhatla, I. Stewart, R. Mihalcea, Surfacing racial stereotypes through identity portrayal, in: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022, ACM, 2022, pp. 1604–1615. URL: https://doi.org/10.1145/3531146.3533217. doi:10.1145/3531146.3533217.

[12] S. Asdaq, S. Alshehri, S. Alajlan, A. Almutiri, A. Alanazi, Depression in persons with disabilities: a scoping review, Front. Public Health (2024). doi:10.3389/fpubh.2024.1383078.

[13] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proc. Natl. Acad. Sci. USA 115 (2018) E3635–E3644. URL: https://doi.org/10.1073/pnas.1720347115. doi:10.1073/PNAS.1720347115.

[14] S. Kiritchenko, S. M. Mohammad, Examining gender and race bias in two hundred sentiment analysis systems, in: M. Nissim, J. Berant, A. Lenci (Eds.), Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018, Association for Computational Linguistics, 2018, pp. 43–53. URL: https://doi.org/10.18653/v1/s18-2005. doi:10.18653/V1/S18-2005.

[15] E. Sheng, K. Chang, P. Natarajan, N. Peng, Societal biases in language generation: Progress and challenges, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meet-

ing of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, Association for Computational Linguistics, 2021, pp. 4275–4293. URL: https://doi.org/10.18653/v1/2021.acl-long.330. doi:10.18653/V1/2021.ACL-LONG.330.

[16] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, S. Denuyl, Social biases in NLP models as barriers for persons with disabilities, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 5491–5501. URL: https://doi.org/10.18653/v1/2020.acl-main.487. doi:10.18653/V1/2020.ACL-MAIN.487.

[17] K. Mei, S. Fereidooni, A. Caliskan, Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023, ACM, 2023, pp. 1699–1710. URL: https://doi.org/10.1145/3593013.3594109. doi:10.1145/3593013.3594109.

[18] A. Salinas, P. Shah, Y. Huang, R. McCormack, F. Morstatter, The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama, in: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3617694.3623257. doi:10.1145/3617694.3623257.

[19] E. M. Smith, M. Hall, M. Kambadur, E. Presani, A. Williams, "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9180–9211. URL: https://aclanthology.org/2022.emnlp-main.625/. doi:10.18653/v1/2022.emnlp-main.625.

[20] W. H. Organization, World Health Organization - Disability, https://www.who.int/health-topics/disability, 2023. Accessed: 2025-01-13.

[21] P. N. Venkit, M. Srinath, S. Wilson, A study of implicit bias in pretrained language models against people with disabilities, in: N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, International Committee on Computational Linguistics, 2022, pp. 1324–1332.

[22] Z. Chu, Z. Wang, W. Zhang, Fairness in large language models: A taxonomic survey, SIGKDD Explor. Newsl. 26 (2024) 34–48. URL: https://doi.org/10.1145/3682112.3682117. doi:10.1145/3682112.3682117.

[23] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, S. R. Bowman, BBQ: A hand-built bias benchmark for question answering, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 2086–2105. URL: https://doi.org/10.18653/v1/2022.findings-acl.165. doi:10.18653/V1/2022.FINDINGS-ACL.165.

[24] R. Qian, C. Ross, J. Fernandes, E. M. Smith, D. Kiela, A. Williams, Perturbation augmentation for fairer NLP, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9496–9521. URL: https://aclanthology.org/2022.emnlp-main.646/. doi:10.18653/v1/2022.emnlp-main.646.

[25] N. Tilmes, Disability, fairness, and algorithmic bias in AI recruitment, Ethics Inf. Technol. 24 (2022) 21. URL: https://doi.org/10.1007/s10676-022-09633-2. doi:10.1007/S10676-022-09633-2.

[26] K. S. Glazko, Y. Mohammed, B. Kosa, V. Potluri, J. Mankoff, Identifying and improving disability bias in gpt-based resume screening, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024, ACM, 2024, pp. 687–700. URL: https://doi.org/10.1145/3630106.3658933. doi:10.1145/3630106.3658933.

[27] M. Díaz, I. Johnson, A. Lazar, A. M. Piper, D. Gergle, Addressing age-related bias in sentiment analysis, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 6146–6150. URL: https://doi.org/10.24963/ijcai.2019/852. doi:10.24963/IJCAI.2019/852.

[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Con-

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI (2019).

[30] S. Hassan, M. Huenerfauth, C. O. Alm, Unpacking the interdependent systems of discrimination: Ableist bias in NLP systems through an intersectional lens, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, Association for Computational Linguistics, 2021, pp. 3116–3123. URL: https://doi.org/10.18653/v1/2021.findings-emnlp.267. doi:10.18653/V1/2021.FINDINGS-EMNLP.267.

[31] B. Herold, J. Waller, R. Kushalnagar, Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies, in: S. Ebling, E. Prud'hommeaux, P. Vaidyanathan (Eds.), Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 58–65. URL: https://aclanthology.org/2022.slpat-1.8/. doi:10.18653/v1/2022.slpat-1.8.

[32] R. Li, A. Kamaraj, J. Ma, S. Ebling, Decoding ableism in large language models: An intersectional approach, in: D. Dementieva, O. Ignat, Z. Jin, R. Mihalcea, G. Piatti, J. Tetreault, S. Wilson, J. Zhao (Eds.), Proceedings of the Third Workshop on NLP for Positive Impact, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 232–249. URL: https://aclanthology.org/2024.nlp4pi-1.22/. doi:10.18653/v1/2024.nlp4pi-1.22.

[33] G. V. Aher, R. I. Arriaga, A. T. Kalai, Using large language models to simulate multiple humans and replicate human subject studies, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 337–371. URL: https://proceedings.mlr.press/v202/aher23a.html.

[34] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, D. Wingate, Out of one, many: Using language models to simulate human samples, Political Analysis 31 (2023) 337–351. doi:10.1017/pan.2023.2.

[35] G. Gui, O. Toubia, The challenge of using llms to simulate human behavior: A causal inference perspective, CoRR abs/2312.15524 (2023). URL: https://doi.org/10.48550/arXiv.2312.15524. doi:10.48550/ARXIV.2312.15524. arXiv:2312.15524.

[36] T. Hu, N. Collier, Quantifying the persona effect in LLM simulations, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 10289–10307. URL: https://doi.org/10.18653/v1/2024.acl-long.554. doi:10.18653/V1/2024.ACL-LONG.554.

[37] W. Li, J. Liu, A. Liu, X. Zhou, M. Diab, M. Sap, BIG5-CHAT: shaping LLM personalities through training on human-grounded data, CoRR abs/2410.16491 (2024). URL: https://doi.org/10.48550/arXiv.2410.16491. doi:10.48550/ARXIV.2410.16491.

[38] E. Frantar, S. Ashkboos, T. Hoefler, D. Alistarh, GPTQ: accurate post-training quantization for generative pre-trained transformers, CoRR abs/2210.17323 (2022). URL: https://doi.org/10.48550/arXiv.2210.17323. doi:10.48550/ARXIV.2210.17323.

[39] J. J. Al-Menayes, Motivations for using social media: An exploratory factor analysis, International Journal of Psychological Studies 7 (2015) 43.

[40] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977).

[41] C. J. Hutto, E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, A. Oh (Eds.), Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014, The AAAI Press, 2014.

[42] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Comput. Intell. 29 (2013) 436–465. URL: https://doi.org/10.1111/j.1467-8640.2012.00460.x.

[43] Y. Li, J. Chan, G. Peko, D. Sundaram, Mixed emotion extraction analysis and visualisation of social media text, Data Knowl. Eng. 148 (2023) 102220. URL: https://doi.org/10.1016/j.datak.2023.102220. doi:10.1016/J.DATAK.2023.102220.

[44] R. Poświata, M. Perełkiewicz, OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 276–282. URL: https://aclanthology.org/2022.ltedi-1.40. doi:10.18653/v1/2022.ltedi-1.40.

[45] Z. Wyatt, The dark side of #positivevibes: Understanding toxic positivity in modern culture, Psychiatry and Behavioral Health 3 (2024) 1–6.