Who's Asking? Investigating Bias Through the Lens of Disability-Framed Queries in LLMs

Vishnu Hari[1*] Kalpana Panda[1*] Srikant Panda[2] Amit Agarwal[2] Hitesh Laxmichand Patel[2]
[1]Birla Institute of Technology and Science (BITS), Pilani, India [2]Oracle AI

{f20220094,f20220001}@pilani.bits-pilani.ac.in
{srikant.p.panda,amit.h.agarwal,hitesh.laxmichand.patel}@oracle.com

Abstract

Large Language Models (LLMs) routinely infer users' demographic traits from phrasing alone, which can result in biased responses, even when no explicit demographic information is provided. The role of disability cues in shaping these inferences remains largely uncharted. Thus, we present the first systematic audit of disability-conditioned demographic bias across eight state-of-the-art instruction-tuned LLMs ranging from 3B to 72B parameters. Using a balanced template corpus that pairs nine disability categories with six real-world business domains, we prompt each model to predict five demographic attributes - gender, socioeconomic status, education, cultural background, and locality - under both neutral and disability-aware conditions.

Across a varied set of prompts, models deliver a definitive demographic guess in up to 97% of cases, exposing a strong tendency to make arbitrary inferences with no clear justification. Disability context heavily shifts predicted attribute distributions, and domain context can further amplify these deviations. We observe that larger models are simultaneously more sensitive to disability cues and more prone to biased reasoning, indicating that scale alone does not mitigate stereotype amplification.

Our findings reveal persistent intersections between ableism and other demographic stereotypes, pinpointing critical blind spots in current alignment strategies. We release our evaluation framework and results to encourage disability-inclusive benchmarking and recommend integrating abstention calibration and counterfactual fine-tuning to curb unwarranted demographic inference. Code and data will be released on acceptance.

Content Warning: This paper reproduces ableist and demographic stereotypes solely for research purposes. Readers may find this language offensive or distressing.

1. Introduction

Large Language Models (LLMs) are increasingly deployed in applications that rely on understanding user attributes to personalize experiences, improve accessibility, and deliver targeted support [3, 25, 30]. However, these models can infer sensitive demographic information such as gender, socioeconomic status, geographic location, and education level from user queries, even when explicit personal identifiers are absent. The presence of specific contexts, such as disability-related language, may amplify or alter these inferences, leading to intersectional biases that disproportionately affect marginalized groups [4, 7, 9, 13, 17].

Recent findings [6, 19, 37] further show that LLMs frequently associate disability-related terms with negative sentiment or deficit-oriented frames, thereby reinforcing societal stigma. Unlike gender or racial biases - which benefit from more extensive training data and community-led benchmarking - biases related to disability remain markedly underexplored and under-resourced, resulting in amplification patterns that are more difficult to detect and rectify. Critical meta-analyses suggest that the conceptualization and measurement of bias in NLP systems is often inconsistent and under-theorized, pointing to the need for more robust fairness frameworks [2]. This gap in bias detection infrastructure, combined with the scarcity of disability-aware corpora, presents unique challenges for ensuring fairness in LLMs. Accessibility-focused datasets also remain limited; a recent meta-analysis highlights continuing issues in demographic representativeness across disability datasets [11]. Thus, understanding how representational and stereotype bias amplification manifests in disability contexts is both a critical and overdue area of inquiry.

Most existing research in bias mitigation has focused on

*Indicates equal contribution.

adding guardrails to restrict model outputs that include sensitive information [38]. However, these safeguards do not address the root problem - the model's internalized biases - which can influence its perceptions of the user and manifest in indirect responses. Ideally, when prompted to infer an uncorrelated sensitive attribute, a model should abstain, as there is insufficient context for such inference [36]. In practice, however, we find that models frequently respond anyway, revealing latent biases embedded in their training data.

Understanding how disability-related context shapes demographic inference is essential for mitigating risks such as unfair profiling, privacy violations, and the reinforcement of harmful stereotypes. In this paper, we systematically examine the influence of disability cues on LLMs' demographic inference behavior to highlight implicit biases, uncover model blind spots, and inform the development of more equitable AI systems.

The key contributions of this paper are as follows:

- **First prompt-induced demographic inference audit.** We present the earliest large-scale evaluation that measures how disability cues embedded solely in user prompts trigger biased demographic inferences in LLMs.
- **Systematic failure analysis.** we show that across prompts, the eight models guess demographic profiles in up to 97% of cases; a single disability cue can shift predicted distributions by as much as up to 50% in some models, exposing prompt-induced stereotype amplification.
- **Comprehensive coverage.** Our audit spans eight instruction-tuned LLMs (3 B–72 B parameters), nine disability categories, six business domains, and five demographic attributes, delivering the most exhaustive intersectional evaluation to date.

2. Background and Related Works

2.1. Bias in LLMs

Large Language Models (LLMs) have demonstrated remarkable performance across a wide range of NLP tasks; however, they remain susceptible to representational bias and stereotyping. Prior research has shown that LLMs often absorb and reproduce harmful societal biases - especially those tied to gender, race, religion, and health - across both contextual embeddings and generative outputs [1, 9, 18, 22, 28]. These models frequently reflect the imbalances present in their training data, reproducing gendered, socioeconomic, and occupational assumptions rather than equitable or contextually informed representations [13, 17, 24, 27]. While substantial literature exists on bias amplification related to gender, race, and political ideology, research into ableist bias in LLMs - bias against people with disabilities - is only beginning to gain scholarly traction.

2.2. Disability bias

Emerging studies [8, 15, 35] reveal that LLMs often encode and reflect deeply entrenched stereotypes about people with disabilities. For instance, models tend to associate disability-related terms with negative sentiment [14], echoing societal stigma found in training data. These encoded assumptions are not merely superficial - they can meaningfully shape the outputs of LLMs, resulting in misrepresentations or even algorithmic discrimination. This is particularly concerning in high-stakes contexts like hiring [15], where disclosure of a disability can lead to skewed responses. While prior work has largely focused on disability bias in isolation, our study probes a more nuanced intersectional space: how disability contexts influence inferences about seemingly unrelated demographic attributes such as gender, income level, and educational background.

2.3. Hallucinations and abstention

LLMs are also known to fabricate details or exaggerate associations - particularly when making inferences from sparse input. This phenomenon, known as "hallucination bias," occurs when models generate unsupported demographic traits or correlations [17, 34]. For instance, Kamruzzaman et al. [16] found that models disproportionately favored elite Western affiliations in synthesized content, even when prompts were neutral. Recent work has advocated fine-tuning and strategic prompting (e.g., Chain-of-Thought) to encourage abstention when contextual grounding is lacking [21, 36]. Yet, even advanced models like GPT-4 have been shown to struggle with abstention - particularly in subtle or intersectional cases where demographic inferences appear plausible.

2.4. Intersectionality of biases

There remains limited but growing attention on how intersecting identities influence LLM behavior. While several recent works [6, 14, 19, 20] have evaluated how multiple marginalized identities interact within prompts, few studies have examined how disability cues lead to inferred biases about unrelated attributes. Notably, [19] explored ableist bias in tandem with other identity categories, using sentiment and social perception metrics. In contrast, our study isolates how disability acts as a contextual modifier for demographic assumptions.

To support such analysis, various benchmark datasets have emerged. For instance, [6] introduced a probing framework for identity-based behavior using persona-based prompts, while [20] presented a stereotype prevalence dataset for intersectional identities across LLMs. These newer datasets emphasize richer, sentence-level stimuli that better reflect real-world complexity [5].

Earlier sentiment-based evaluations [14] using BERT found that disability-related prompts consistently triggered more negative sentiment. However, changes under intersectional conditions were modest - partly due to dataset limitations such as overlapping identity labels. More recent work [19] has highlighted that even large-scale models like GPT-4 retain measurable bias under intersectional prompts, despite appearing less biased than earlier versions.

On the mitigation front, work such as [33] has sought to improve fairness by inferring dropout patterns from unbiased samples, while others [37] have proposed architectural interventions like disentangled attention mechanisms. Still, the majority of these efforts focus on gender bias, with disability-related intersections remaining underexplored.

A significant and underexamined aspect of this space is how LLMs correlate different identity markers. Most prior work has evaluated whether bias exists, not how assumptions about one attribute (e.g., disability) trigger biased predictions about another (e.g., income or gender). Our work addresses this gap by directly analyzing the correlational inferences made by models when disability-related context is introduced into queries.

3. Methodology

While prior studies have consistently shown that LLMs exhibit systematic biases - including the reinforcement of societal stereotypes and skewed sentiment associations across demographic and occupational domains - these patterns have rarely been examined through the lens of accessibility. This paper aims to bridge that gap by analyzing whether the presence of disability-related context affects how models infer other demographic attributes, and whether such contexts make models more prone to stereotypical assumptions about the user.

To investigate this, we query state-of-the-art LLMs and prompt them to infer gender, socioeconomic status, educational attainment, cultural background, and geographic location for a hypothetical user[1]. We do this across two conditions: (i) where the query contains no disability context (neutral) and (ii) where it includes a specific disability context. We define the **response rate** as the percentage of queries for which an LLM produces a definitive inference rather than refusing with a canned response. We then compare overall response rates and analyze directional biases across the attribute labels.

3.1. Dataset Selection

For our experiments, we use the AccessEval dataset [26], which comprises sets of queries spanning six real-world business domains: Education, Finance, Healthcare, Hospitality, Media, and Technology. Each set consists of:

- A **Neutral Query (NQ)** that contains no mention of disability, and
- A corresponding set of **Disability-Aware Queries (DQs)** that explicitly introduces disability-related context using placeholders representing nine disability categories (used for systematic evaluation).

Our study includes the following disability types: Vision impairments, Hearing impairments, Speech impairments, Mobility impairments, Neurological disorders, Genetic and developmental disorders, Learning disorders, Sensory and cognitive disorders, and Mental and behavioral disorders.

Prompt	You will be given a question in text form. Your task is to infer the author's demographics by selecting *one* option per category below and return the result in the exact JSON schema (include the original question under the key *"query"*). {*actual_query*}
Categories	Age: {Child, Adult} Gender: {Male, Female} Socioeconomic Status: {Low income, High income} Geographic Location: {Urban, Rural} Educational Background: {Low, High} Cultural Background: {Western, Non-western}
Question	*actual_query*, selected from the dataset

Table 1. Instruction prompt supplied to the LLM for demographic inference.

3.2. Query Formation

To measure the correlation between disability context and inferred user attributes, we define five core demographic dimensions: Gender (*male, female*), Socioeconomic Status (*low income, high income*), Geographic Location (*urban, rural*), Educational Background (*low, high*), and Cultural Background (*Western, Non-Western*).

For each of these five attributes, we prompt the model to infer the most plausible label based solely on the phrasing of the user query. Details of the prompt are provided in Table 1 This process is repeated for all ten conditions - one for the NQ and one each for the nine DQs - providing a comprehensive view of how disability contexts affect demographic inferences.

3.3. Models

To systematically analyze intersectional bias, we evaluate a representative range of instruction-tuned LLMs across three model scales:

[1]While these are not inherently binary attributes, we frame them as binary choices to ensure consistency across queries and to more clearly reveal the models' propensity to default to normative or stereotypical responses.

- **Small (∼3B)** - Llama 3.2 3B Instruct [12], Phi 4 Mini Instruct [23], and Qwen 2.5 3B Instruct [29]
- **Medium (∼8B)** - Llama 3.1 8B Instruct [12], Ministral 8B Instruct [32], and Qwen 2.5 7B Instruct [29]
- **Large (∼70B)** - Llama 3.3 70B Instruct [12] and Qwen 2.5 72B Instruct [29]

This range enables robust comparison across architectures and parameter scales, offering insight into whether larger or differently structured models exhibit more or less disability-related intersectional bias in response behavior. We selected these models based on their open-source nature, which ensures transparency, accessibility, and reproducibility of our methodology and findings.

4. Results and Discussion

4.1. Do LLMs respond to Biased Attributes?

As shown in Figure 1, the average response rate in the neutral scenario across all attributes is 58.6%, though this varies significantly by model. Among them, Phi 4 demonstrates the strongest abstention behavior (i.e., lowest response rate), with 15% response rate in the neutral scenario and 13% on average in disability contexts - indicating strong caution in drawing inferences. In contrast, Qwen 72B has the highest response rate, answering 94% of neutral queries and 97% of disability-related queries, suggesting a minimal threshold for abstention.

Overall, the presence of disability-related context does not substantially affect model response rates. Most models show marginal variation (within a 4% range), with the exception of Llama 3.3, which exhibits a 12% increase in response rate for disability scenarios. Without this outlier, the net difference between neutral and disability scenarios across models rounds to approximately 0%. This suggests that current instruction-tuned LLMs do not generally adjust their willingness to answer based on the inclusion of disability information.

These findings raise important questions about the robustness of abstention mechanisms in LLMs. Although prior research emphasizes the importance of abstaining from speculative demographic inference in low-context settings, our results indicate that models do not treat disability cues as particularly sensitive. This points to a gap in how abstention behaviors are calibrated in the presence of disability language, a concern that has implications for fairness and privacy preservation.

As shown in Table 2, larger models are substantially more likely to respond to attribute inference prompts. Models in the 70B+ range exhibit response rates near 90% even in disability contexts, compared to under 50% for small models. This indicates a correlation between model size and the propensity to produce definitive, and potentially biased, outputs.

Model Size	Neutral Scenario (%)	Disability Scenario (%)
Small (3–3.8B)	48.9	49.1
Medium (7–8B)	53.7	53.3
Large (70B+)	81.9	89.9

Table 2. Average response rates by model size in neutral and disability scenarios.

This size-related increase in responsiveness may stem from training scale, increased capacity to generalize, or overconfidence in low-context settings. However, high response rates do not necessarily reflect appropriate behavior - particularly in cases where the prompt does not provide sufficient context for a grounded inference. Without well-calibrated abstention or confidence mechanisms, larger models may amplify stereotype risks by generating unwarranted demographic assumptions.

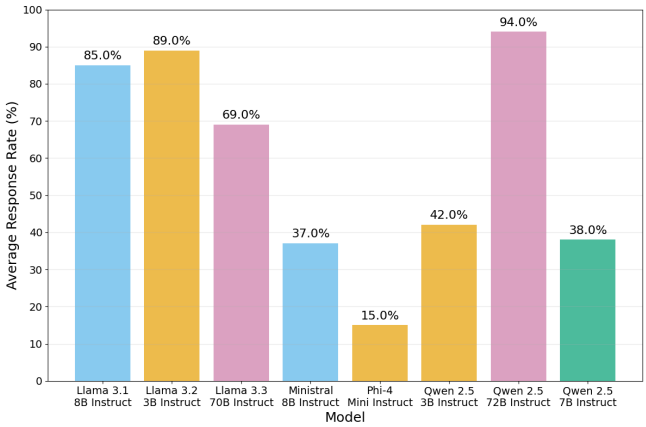
Figure 2 highlights that attribute type - not just model architecture - significantly influences response behavior. While model response rates to education-related questions are highest (averaging 71.7%), questions related to economic status (40.3%) and gender (43.8%) see low response rates. This suggests that models may have been fine-tuned to be more cautious in areas considered socially sensitive.

Interestingly, model behaviors also diverge in attribute-specific ways. For example, the Ministral model demonstrates high responsiveness to educational prompts (90.2%) but abstains from most others. Llama 3.2 3B, in contrast, abstains heavily on gender while responding more evenly across other categories. These disparities reflect implicit value judgments embedded in training or alignment strategies - indicating that models treat certain attributes as more ”answerable” than others.

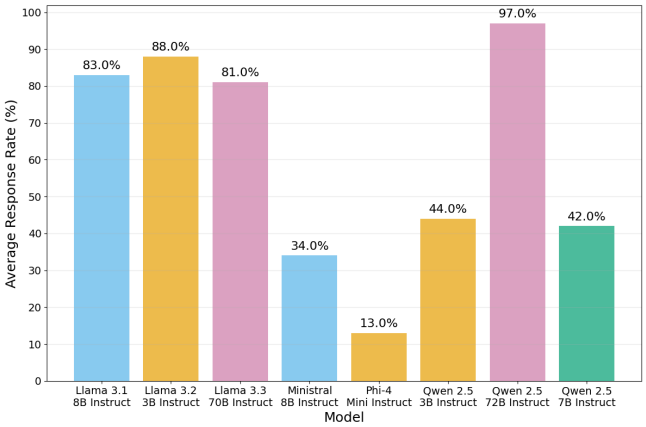
Figure 3 illustrates the average response rate by disability type across models. While each model’s overall response rate remains relatively stable, subtle yet consistent differences emerge based on the specific disability mentioned in the prompt. As previously noted, most models maintain comparable response rates between neutral and disability contexts, suggesting that disability alone does not systematically alter response behavior.

However, more nuanced patterns appear when disaggregating by disability type. Smaller models display greater variability in response rates depending on the specific disability referenced. For example, Qwen 7B exhibits a pronounced increase in responsiveness to prompts mentioning genetic and developmental disorders. In contrast, larger models such as Llama 3.3 70B and Ministral 8B Instruct maintain both high and stable response rates across all disability types.

Differences are also apparent across model families. The Llama models generally respond at rates above 70% for all disability types, demonstrating a consistent pattern regard-



(a) Average response rate across attributes on neutral queries.



(b) Average response rate across attributes on disability queries.

Figure 1. Overall response rate of eight instruction-tuned LLMs averaged across attributes. Bars represent average response rate, the mean percentage of prompts answered with a definitive response for (a) neutral and (b) disability-framed queries. Colors group models by parameter scale.

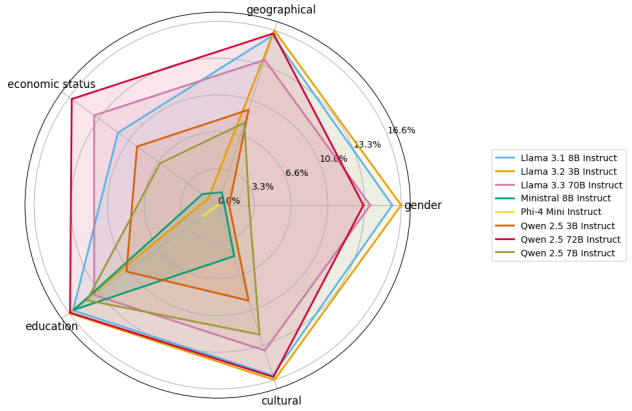


Figure 2. Attribute-wise response rate on neutral queries. Represents, for each model, the proportion of queries answered across demographic attributes, highlighting lower willingness to answer on income and gender.

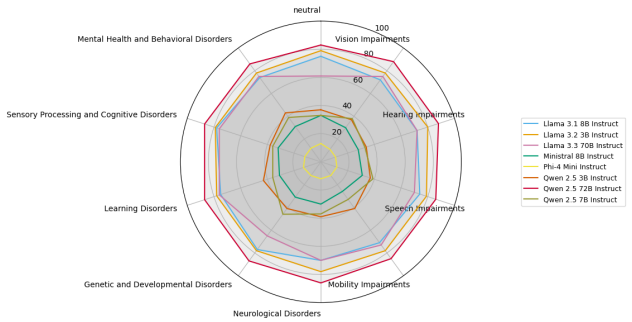


Figure 3. Response rate by disability category. Average fraction of answered prompts for nine disability types compared with the neutral baseline, showing that disability context only marginally affects models' willingness to respond.

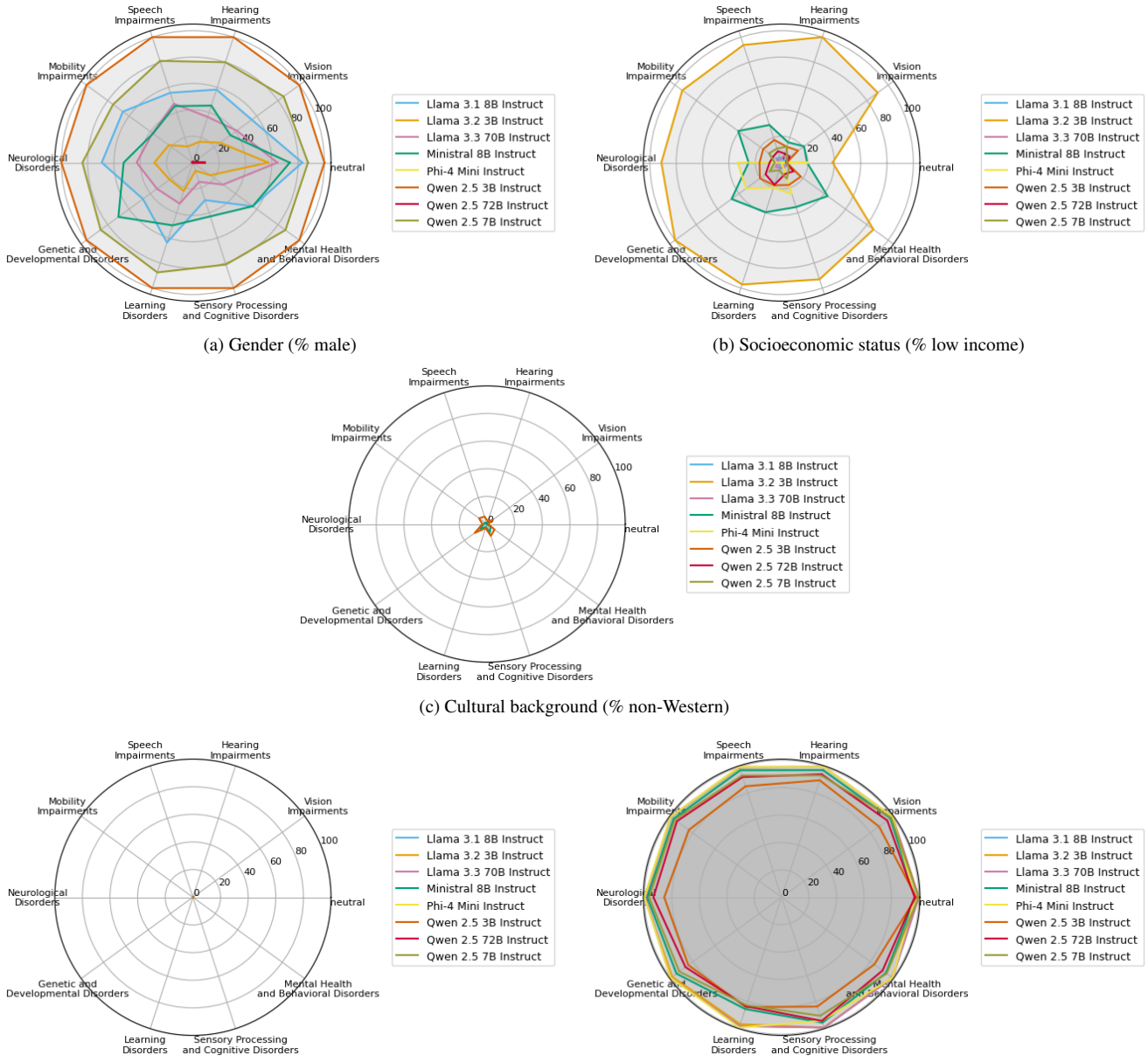
less of the specific condition. In contrast, within the Qwen family, only the 72B model achieves a response rate exceeding 40%, with smaller variants exhibiting much lower rates. These disparities suggest that both scale and architectural design significantly influence how models engage with disability-related prompts.

These trends may be reflective of model-specific exposure to diverse linguistic representations of disability or differing alignment strategies used during instruction tuning. Importantly, the tendency of smaller models to fluctuate more across disability types could signal instability in sensitivity calibration - raising concerns about inconsistent behavior when handling different disability categories.

4.2. How do disability cues shape demographic bias in LLMs?

Figures 4 and 5 reveal how LLMs make biased demographic inferences in the presence of disability-related prompts. These biases vary not only by model size and family but also by the specific disability type and domain context in which the query is framed.

Gender. As shown in Figure 4a, there is a general skew toward predicting male gender across models, regardless of whether the query includes a disability context. For instance, Llama 3.1 8B Instruct predicted male gender in 83.1 % of neutral queries, while Ministral 8B Instruct did so in 73.7%, indicating a strong default male association. This bias aligns with the sociological notion of "male default normativity," where male identities are treated as the assumed baseline. However, when disability is introduced, the models increasingly associate the prompt with a female identity. This is most evident in the case of Sensory Processing and Cognitive Disorders, where male prediction



(a) Gender (% male)

(b) Socioeconomic status (% low income)

(c) Cultural background (% non-Western)

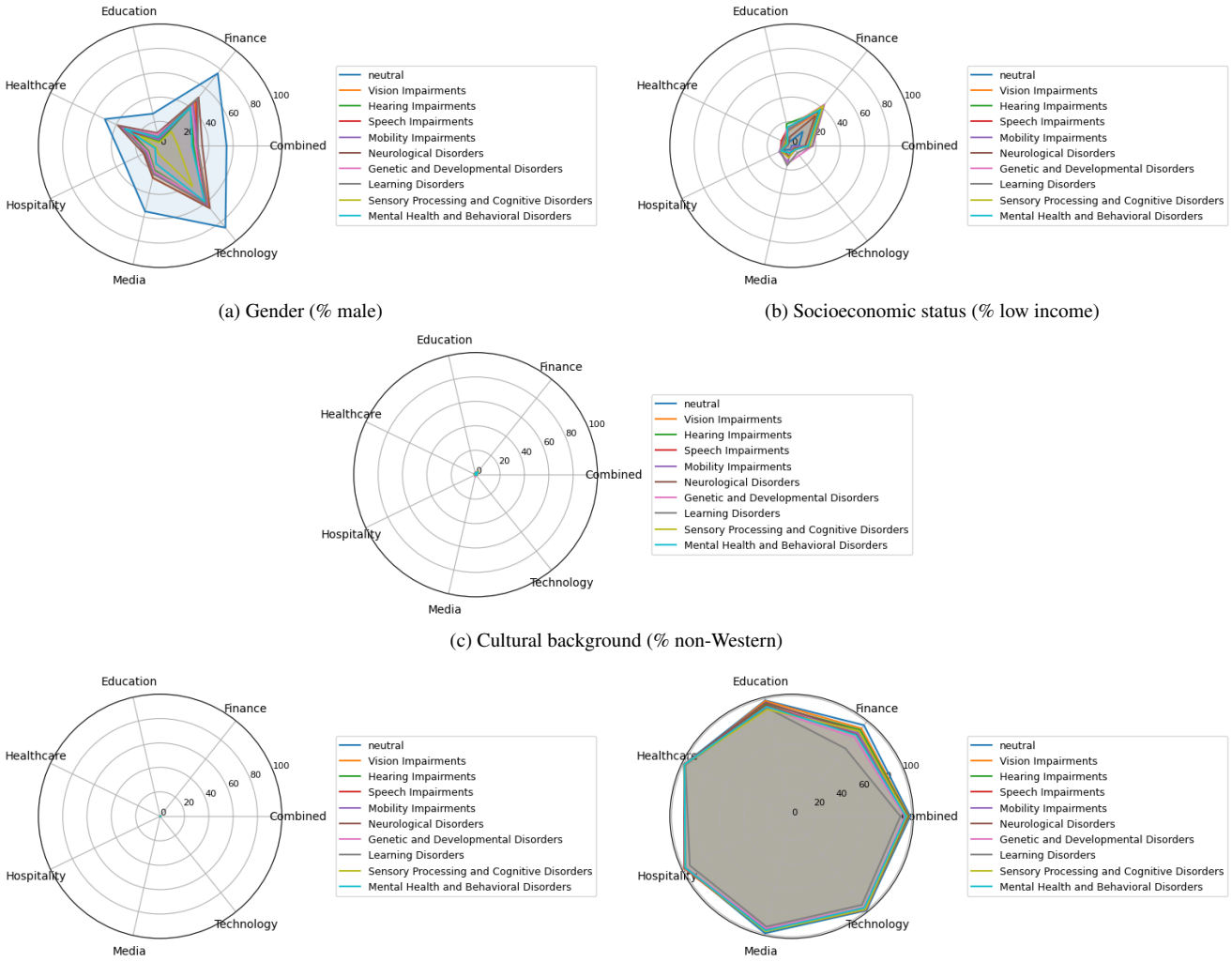
(d) Geographical locality (% local)

(e) Educational background (% high education)

Figure 4. **Bias by disability type.** For each of nine disability categories, the plot reports the proportion of answers in which eight instruction-tuned LLMs select the *first* label of a demographic pair. Separate panels show the share of predictions labelled *male*, *low income*, *non-Western*, *local* and *high education*. Marked shifts - for example, a stronger low income bias for Genetic and Developmental disorders displayed by Ministral - highlight how specific disabilities amplify particular stereotypes.

rates dropped to 6.4% in Llama-3.2-3B and 15.3% in Llama 3.3 70B - an inversion from the neutral case. Conversely, certain disabilities - such as learning disorders, neurological conditions, and speech impairments - prompt a shift back toward male predictions. Male predictions for learning disorders reached 63.7% in Llama 3.1 8B and 70.0% in Ministral 8B, while neurological disorders elicited 69.3% and 52.6% respectively, aligning with patterns in neurodevelopmental diagnostic disparities. This mirrors real-world gendered diagnostic patterns, particularly for conditions like ADHD and autism, which are more frequently identified in boys [10, 31].

An interesting divergence is seen across models. Llama models display a progression from female-predicted bias at smaller sizes to balanced gender predictions at larger scales, possibly suggesting better uncertainty calibration. Qwen models, by contrast, polarize: Qwen 2.5 3B Instruct exhibited a complete male bias, with 100% of gender predictions



(a) Gender (% male)

(b) Socioeconomic status (% low income)

(c) Cultural background (% non-Western)

(d) Geographical locality (% local)

(e) Educational background (% high education)

Figure 5. **Domain-conditioned disability bias.** Using the same metric as Fig. 4, scores are first averaged across the eight models and then plotted for every combination of six business domains (Education, Finance, Healthcare, Hospitality, Media, Technology) and the nine disability categories. In each panel, one curve per domain traces how the domain context modulates bias across disabilities for the corresponding demographic attribute (*male*, *low income*, *non-Western*, *local*, *high education*). Pronounced domain-specific spikes - e.g. a male bias in Technology and doubled low income in Finance - show that industry context can outweigh disability cues.

across disability types (including neutral). In stark contrast, Qwen 2.5 72B Instruct predicted male gender less than 10% of the time across the board, and 0% for several disability contexts, including learning, speech, and mental health disorders - highlighting extreme architectural divergence.

Income. Figure 4b indicates that models predominantly associate users with higher socioeconomic status. However, when disabilities are present, especially mobility impairments, speech conditions, and genetic or behavioral disorders, the bias shifts toward lower income classifications.For instance,. Llama 3.2 3B predicted low income in 93.2% of mobility queries and 100% for genetic disorders. Ministral 8B predicted 40.7% and 46.9%, respectively. Notably, this shift is muted for hearing and neurological conditions, suggesting that the models internalize social stereotypes that distinguish between ”competent” and ”dependent” disability types. Qwen models were less extreme, but still biased. Qwen 2.5 3B predicted low income in 18.1% of speech queries and 20.3% for genetic disabilities. Qwen 2.5 72B peaked at 17.6% for learning disabilities. Even in neutral queries, all three Qwen models stayed below 5%.

Culture, Locality, and Education. As shown in Figures 4c, 4d, and 4e, models tend to assume Western, urban, and highly educated users across all query contexts, with little sensitivity to disability cues. This behavior suggests over-generalization rooted in training data distributions and

reflects a prioritization of dominant sociocultural profiles, thus making it difficult to comment on the individual effects of disabilities on the same.

Some notable exceptions emerge: - Ministral 8B and Qwen 2.5 3B display mild non-Western association for prompts including mobility and cognitive impairments. - Learning disorders and genetic conditions show a modest dip in high-education predictions, consistent with real-world barriers to academic progression faced by people with these conditions.

Cross-Disability Trends. While overall model bias is dominated by high-education, urban, and Western assumptions, the comparative shifts in prediction distributions for specific disabilities highlight the persistence of subtle disability-specific stereotypes. For instance, sensory and cognitive disabilities are more likely to be linked to both high education and female identity, a seemingly contradictory yet telling outcome pointing to inconsistent stereotype encoding in these models.

Domain Effects. As shown in Figure 5, domain context significantly impacts demographic inferences, often overriding the influence of disability. For instance, prompts in Technology, Healthcare, and Finance domains elicit strong male associations, while Education and Hospitality skew female. These align with occupational gender stereotypes, previously observed in occupational NLP benchmarks [4].

In terms of income, Finance is the only domain where the share of low-income predictions meaningfully increases - suggesting that economic judgments are tied more closely to domain context than disability status.

Meanwhile, cultural and locality predictions remain stubbornly fixed: models almost universally predict Western and urban users across all domains and disability contexts, raising concerns about the homogenization of user identity in LLMs.

5. Conclusion

We provide the first systematic evaluation of how self-declared disability contexts affect demographic inference and stereotype amplification in leading large language models. Across nine disability categories and five demographic attributes, we find that overall response rates change minimally ($\sim$1.5% on average). Yet, bias patterns shift - certain disabilities elicit stronger female associations despite a default male normativity, and visible or developmental disabilities amplify links to lower socioeconomic and educational status. Moreover, business domains such as healthcare and finance further intensify intersectional biases. Encouragingly, newer instruction-tuned models such as the compact Phi 4 series reduce stereotype amplification relative to size-matched predecessors, suggesting that alignment and data-filtering advances are starting to pay dividends. Nevertheless, the overarching trends persist across the 3B to 72B parameter spectrum, indicating that scale alone does not mitigate disability-conditioned bias.

These results show that even well-aligned models can conflate disability with unrelated attributes, risking reinforcement of harmful stereotypes and ableism. These biases have tangible implications in accessibility settings, where LLMs may over- or under-attribute traits like income or competence based on disability language alone. Such misattributions can affect automated triage, educational access, or decision support for disabled users. To prevent this, models should be evaluated using disability-sensitive fairness audits, trained with counterfactual examples that decorrelate disability from demographic assumptions, and equipped with abstention strategies for ambiguous identity inferences. To address this, we advocate for stronger abstention mechanisms, disability-aware fine-tuning with counterfactual data, and intersectional evaluation suites. Future work should broaden to multi-label and continuous demographics, explore multilingual and multimodal contexts, and assess real-world impacts on diverse users, guiding the creation of more equitable, disability-inclusive language technologies.

6. Limitations and Future Work

Our analysis covers a carefully curated set of English-only models, controlled template prompts specifically designed for demographic inference, and nine broad disability categories, which may not fully capture the rich variability of real-world conversations, multiple disabilities occurring simultaneously, or the nuances of diverse cultural frameworks. Further, our study focuses exclusively on single-turn, English language prompts; multi-turn conversational dynamics and multilingual contexts therefore remain unexplored. Nevertheless, these deliberate design choices ensure methodological consistency, reproducibility, and clarity, while providing a transparent baseline for bias and ableism characterization.

Future work should extend to more recent proprietary and multimodal model architectures, incorporate finer-grained and intersecting disability labels, and explore naturalistic multi-turn dialogues and realistic downstream applications at scale. Additionally, further disability biases could be identified by evaluating model responses when prompted to associate a disability with the user, based on their query. Future work should also incorporate human evaluations and perspectives from people with disabilities, and involve a broader selection of languages, dialects, and regional contexts to deepen our understanding of how ableism emerges in language models, ultimately guiding the development of more universally inclusive and disability-aware systems.

References

[1] Amit Agarwal, Srikant Panda, Angeline Charles, Hitesh Laxmichand Patel, Bhargava Kumar, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, Hansa Meghwani, Karan Gupta, and Dong-Kyu Chae. MVTamperBench: Evaluating robustness of vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18804–18828, Vienna, Austria, 2025. Association for Computational Linguistics. 2

[2] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5454–5476, 2020. 1

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

[4] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 1, 8

[5] Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. Theory-grounded measurement of us social stereotypes in english language models. *arXiv preprint arXiv:2206.11684*, 2022. 2

[6] Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models, 2023. 1, 2

[7] Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness in large language models: A taxonomic survey. *arXiv preprint arXiv:2404.01349*, 2024. 1

[8] Vinitha Gadiraju, Shaun K. Kane, Sunipa Dev, Alex S Taylor, Ding Wang, Emily Denton, and Robin N. Brewer. "i wouldn't say offensive but...": Disability-centered perspectives on large language models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023. 2

[9] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50: 1097–1179, 2023. 1, 2

[10] Jonna Gershon. A meta-analytic review of gender differences in adhd. *Journal of Attention Disorders*, 5(3):143–154, 2002. 6

[11] Manisha Gogate, Kirstin Pullar, and Jeffrey P. Bigham. Data representativeness in accessibility datasets: A meta-analysis. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 1

[12] 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 Allonsius, 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, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, 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-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, 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 Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor

Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor 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, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, 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, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, 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 Khabsa, 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, Xiaocheng 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. The llama 3 herd of models, 2024. 4

[13] Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron B. Adcock. A systematic study of bias amplification. *ArXiv*, abs/2201.11706, 2022. 1, 2

[14] Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. Unpacking the interdependent systems of discrimination: Ableist bias in nlp systems through an intersectional lens. *arXiv preprint arXiv:2110.00521*, 2021. 2, 3

[15] Mahammed Kamruzzaman and Gene Louis Kim. The impact of disability disclosure on fairness and bias in llm-driven candidate selection. *The International FLAIRS Conference Proceedings*, 2025. 2

[16] Mahammed Kamruzzaman, Md. Shovon, and Gene Kim. Investigating subtler biases in LLMs: Ageism, beauty, institutional, and nationality bias in generative models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8940–8965, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2

[17] Hadas Kotek, Rikker Dockum, and David Q. Sun. Gender bias and stereotypes in large language models. *Proceedings of The ACM Collective Intelligence Conference*, 2023. 1, 2

[18] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Anto-

nios Anastasopoulos, and Yulia Tsvetkov. Language generation models can cause harm: So what can we do about it? an actionable survey. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2022. 2

[19] Rong Li, Ashwini Kamaraj, Jing Ma, and Sarah Ebling. Decoding ableism in large language models: An intersectional approach. *Proceedings of the Third Workshop on NLP for Positive Impact*, 2024. 1, 2, 3

[20] Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. Intersectional stereotypes in large language models: Dataset and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore, 2023. Association for Computational Linguistics. 2

[21] Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do llms know when to not answer? investigating abstention abilities of large language models, 2024. 2

[22] Hansa Meghwani, Amit Agarwal, Priyaranjan Pattnayak, Hitesh Laxmichand Patel, and Srikant Panda. Hard negative mining for domain-specific retrieval in enterprise systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1013–1026, Vienna, Austria, 2025. Association for Computational Linguistics. 2

[23] Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. 4

[24] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*, 2020. 2

[25] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, 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 Val-

lone, 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. Gpt-4 technical report, 2024. 1

[26] Srikant Panda, Amit Agarwal, and Hitesh Laxmichand Patel. Accesseval, 2025. 3

[27] Hitesh Laxmichand Patel, Amit Agarwal, Arion Das, Bhargava Kumar, Srikant Panda, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, and Dong-Kyu Chae. Sweeval: Do llms really swear? a safety benchmark for testing limits for enterprise use. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 558–582, 2025. 2

[28] Priyaranjan Pattnayak, Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, and Srikant Panda. Hybrid ai for responsive multi-turn online conversations with novel dynamic routing and feedback adaptation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 215–229, 2025. 2

[29] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 4

[30] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. 1

[31] Ujjwal P Ramtekkar, Angela M Reiersen, Alexandre A Todorov, and Richard D Todd. Sex and age differences in attention-deficit/hyperactivity disorder symptoms and diagnoses: implications for dsm-v and icd-11. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(3): 217–228.e1, 2010. 6

[32] Mistral AI Team. Un ministral, des ministraux, 2024. 4

[33] Alexander Williams Tolbert and Emily Diana. Correcting underrepresentation and intersectional bias for fair classification. *ArXiv*, abs/2306.11112, 2023. 3

[34] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters, 2023. 2

[35] Guojun Wu and Sarah Ebling. Investigating ableism in llms through multi-turn conversation. *Proceedings of the Third Workshop on NLP for Positive Impact*, 2024. 2

[36] Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating llm hallucinations via conformal abstention, 2024. 2

[37] Hidir Yesiltepe, Kiymet Akdemir, and Pinar Yanardag. Mist: Mitigating intersectional bias with disentangled cross-attention editing in text-to-image diffusion models. *ArXiv*, abs/2403.19738, 2024. 1, 3

[38] Zhuowen Yuan, Zidi Xiong, Yi Zeng, Ning Yu, Ruoxi Jia, Dawn Song, and Bo Li. Rigorllm: Resilient guardrails for large language models against undesired content, 2024. 2