Gender Trouble in Language Models: An Empirical Audit Guided by Gender Performativity Theory

Franziska Sofia Hafner University of Oxford Franziska.Hafner@oii.ox.ac.uk Ana Valdivia University of Oxford Institute of Advanced Studies (UCL) Ana.Valdivia@oii.ox.ac.uk Luc Rocher University of Oxford Luc.Rocher@oii.ox.ac.uk

Abstract

Language models encode and subsequently perpetuate harmful gendered stereotypes. Research has succeeded in mitigating some of these harms, e.g. by dissociating non-gendered terms such as occupations from gendered terms such as 'woman' and 'man'. This approach, however, remains superficial given that associations are only one form of prejudice through which gendered harms arise. Critical scholarship on gender, such as gender performativity theory, emphasizes how harms often arise from the construction of gender itself, such as conflating gender with biological sex. In language models, these issues could lead to the erasure of transgender and gender diverse identities and cause harms in downstream applications, from misgendering users to misdiagnosing patients based on wrong assumptions about their anatomy.

For FAccT research on gendered harms to go beyond superficial linguistic associations, we advocate for a broader definition of 'gender bias' in language models. We operationalize insights on the construction of gender through language from gender studies literature and then empirically test how 16 language models of different architectures, training datasets, and model sizes encode gender. We find that language models tend to encode gender as a binary category tied to biological sex, and that gendered terms that do not neatly fall into one of these binary categories are erased and pathologized. Finally, we show that larger models, which achieve better results on performance benchmarks, learn stronger associations between gender and sex, further reinforcing a narrow understanding of gender. Our findings lead us to call for a re-evaluation of how gendered harms in language models are defined and addressed.

Warning: This article discusses the erasure and pathologization of transgender and gender diverse identities in academic literature and in language models.

CCS Concepts

Computing methodologies → Natural language processing;
 Applied computing → Sociology.

Keywords

Language Models, Gender Performativity, Gender Bias, Empirical Evaluations, Technology Audits

This work is licensed under a Creative Commons Attribution 4.0 International License. *FAccT '25, Athens, Greece* © 2025 Copyright held by the owner/author(s). ACM ISBN 979-8-4007-1482-5/2025/06 https://doi.org/10.1145/3715275.3732112

ACM Reference Format:

Franziska Sofia Hafner, Ana Valdivia, and Luc Rocher. 2025. Gender Trouble in Language Models: An Empirical Audit Guided by Gender Performativity Theory. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3715275.3732112

1 Introduction

Language plays a key role in constructing, reinforcing, and subverting gender norms [1]. The English language, for example, constructs gender by assigning the binary categories 'boy' and 'girl' to babies at birth, reinforces gender through phrases such as 'boys will be boys' and 'throw like a girl', and subverts gender through terms not fitting into the 'boy/girl' binary, such as the use of 'they/them' as singular pronouns. The more frequently some gender norm is expressed through language, the more it is normalized [5]. Conversely, gendered experiences not often expressed through language might be perceived as abnormal, unacceptable, or even unimaginable. Language models (LMs) encode these norms from the books, online news articles, and online forums that make up their training data [2, 15, 19, 27, 42]. Encoding human biases and their representation in popular culture, LM-based technologies can then persistently reproduce prejudice against sexual orientation or gender identity.

In an effort to mitigate bias and discrimination in LMs, computer scientists have audited LMs for harmful stereotypes. This research has contributed to understanding and reducing these harms [21]. For example, researchers have used large dictionaries of 'bias probes' and 'prefix templates' to elicit biased sentence completions [40, 41]. Further research on downstream applications, notably in HR and recruitment as well as in clinical and medical applications, showed that LMs consistently stereotype certain races, ethnicities, and genders. Subsequently, new approaches have been developed which target training datasets (e.g. by re-balancing data with pre-defined list of biased word pairs such as 'he/she' and 'king/queen' [28]), learning phase (e.g. by adding an equalizing regularization term to the loss function that produces embeddings with less association between 'he/she' pronouns and occupations and by using reinforcement learning with human feedback), or postprocessing phase (e.g. by setting prompts that help avoid biased sentence completion, or replacing stereotypical keywords) [21].

However, while debiasing efforts reduce easy-to-see forms of bias in simple benchmarks, they fail to address more complex and deeply ingrained associations [33, 46], especially in adversarial settings [35] or in Chain of Thought settings that prompt models to 'think' [39], both circumventing debiasing efforts. Hofmann et al. [23] found that such efforts can successfully reduce negative overt racist stereotypes but exacerbate covert stereotypes, in the form of dialect prejudice. For example, some models are optimized to reduce the association of 'he/she' pronouns or 'black/white' racial categories with occupations. This may not affect cases where gender is inputed from pronouns other than 'he/she', or race from racialised terms other than 'black/white', e.g. through implicit information such as dialect or writing style. Moreover, models optimized according to specific categories such as 'man' and 'woman' risk essentializing complex social identities such as gender into a narrow set of biological attributes or culturally dominant norms, by reinforcing the idea that these categories are fixed, universal, or exhaustive.

Addressing bias and discrimination in LMs requires moving beyond aligning models with pre-existing human preferences, by leveraging decades of critical research from the social sciences on how social characteristics are and should be represented in language [3]. Research in fields such as gender studies, sociolinguistics, feminist studies, and queer studies highlights the importance of language not merely in representing gender, but also in constructing, reinforcing, and subverting gender norms [1]. Critically, they show that gendered harms often arise from the construction of gender itself, instead of from stereotypical linguistic associations. For example, gender and sex are frequently linked narrowly in language (man linked to male, woman linked to female). This strong association can be harmful both to cisgender people (e.g., if the ability to conceive is narrowly tied to womanhood, this is harmful to cisgender women wo do not want to or cannot get pregnant) and to transgender and gender diverse people (e.g., a transgender woman who cannot get pregnant might be excluded from womanhood on this basis). Additionally, by defining classifications of gender (e.g., man and woman), language can contribute to the erasure of gender identities that are not considered in such classifications, such as transgender and gender diverse people. We build on these insights from gender studies, and apply them to the study of LMs.

We derive three main insights on the importance of language from Butler's works on gender performativity and apply them to LMs: (a) *distinction between gender and sex* (LMs should not essentialize gender by conflating it with biological sex), (b) *meaningful embeddings of diverse genders* (LMs should provide meaningful embeddings for all gender identity terms), and (c) *non-stereotypical representations* (LMs should not encode harmful stereotypes about any gender identity terms).

We test how 16 LMs, including versions of Llama, Mistral, RoBERTa, T5, and GPT, construct gender, by measuring the conditional probability they assign in model predictions across different contexts related to sex, gender, and illnesses. Firstly, we find that most of the tested models appear to follow essentializing patterns regarding the relationship between gendered words and sex characteristics (e.g., associating 'man' with testosterone and 'woman' with estrogen). Crucially, we find that gender and sex are conflated more as models get larger. Secondly, some of the models consistently associate unrelated non-human words with sex characteristics more strongly than they do words associated with transgender and gender diverse people, such as the words 'nonbinary', 'transgender', 'genderqueer', 'genderfluid', and 'two-spirit'. This indicates a lack of meaningful embeddings for these words. Finally, most models appear to encode harmful pathologizing associations for transgender and gender diverse words, in particular by associating them with mental illness.

Left unchecked, the essentializing and pathologizing issues we uncovered could have harmful consequences in numerous settings. Our findings affirm the need to consider concepts such as gender beyond simple alignment practices, by engaging critically with them throughout the entire model development process. Ultimately, we hope that efforts to investigate how various social attributes are conceptualized by LMs can lead to the development of models that do not perpetuate, but mitigate, the harms associated with narrow and essentialist definitions.

2 Background

2.1 The Construction of Sex and Gender through Language

Gender and sex are frequently not recognized as separate concepts in discourse on gender. Instead, gender is often framed as binary (every person fits into one of the distinct categories 'man' or 'woman'), immutable (this category can not change), and physiological (this category is assigned based on physical characteristics) [26, p. 88]. This binary, immutable, and physiological understanding of gender is called the 'folk understanding' [26]. While this understanding of gender is not universal across cultures or historical periods, it is dominant in many Western and English-speaking contexts.

The folk understanding of gender has been extensively critiqued in gender studies. In particular, scholars have highlighted how sex and gender are not inherently linked, but become interlinked through their social construction [47]. As Simone de Beauvoir writes in *The Second Sex* on the construction of gender, 'one is not born, but rather becomes, a woman' [11, p. 267]. Similarly, Judith Butler challenges the innateness of sex by asking 'is it natural, anatomical, chromosomal, or hormonal' [5, p. 9]. This untangling of the elements that make up a supposedly binary sex leads Butler to question: 'if the immutable character of sex is contested, perhaps this construct called "sex" is as culturally constructed as gender' [5, p. 9]. Instead of portraying sex as a precursor to gender (as does the folk understanding), Butler therefore argues that gender and sex become interlinked in their social construction [6].

According to gender theorists, language plays a crucial role in the interlinked construction of sex and gender. By describing concepts, language allows us to imagine them more easily. Therefore, the power of language lies in creating an 'imaginable domain' [5, p. 12]. By providing ways to describe some gendered experiences but not others, language can limit which genders and gender expressions are considered normal, acceptable, and even possible. For example, the English language provides many words for articulating binary sex and gender categories (*the male and the man, the female and the woman*). Through this, 'language gains the power to create "the socially real" [5, p. 156].

2.2 The Pathologization of Sex and Gender outside a Binary

When individuals' bodies and gender expressions do not align with the sexed and gendered norms defined through language, individuals face significant social sanctions. Foucault, for example, illustrates this in the book *Herculine Barbin*, where he retraced the struggles of a French intersex person born in 1838. As their body did not fit into a 'male' nor 'female' norm, legal and medical institutions of the time were not able to categorize Barbin within the gender and sex binary [20]. This led to a series of institutional interventions, including medical examinations and legal debates over how to determine someones 'true' sex, hidden behind 'anatomical deceptions' [20, p. 8]. To this day, intersex babies are still routinely operated to fit into a 'male' or 'female' category [31].

In the 20th century, frameworks such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) and International Classification of Diseases (ICD) classified gender variance as 'disordered' [16, 17]. These categorizations both entrenched stigma and positioned medical institutions as gatekeepers of legitimacy and care [9, 13, 24]. While terminology evolved from the diagnosis of a 'gender identity disorder' to 'gender dysphoria' in the 2013 DSM edition (DSM-5), scholars nevertheless argue that the structural underpinnings of gatekeeping persist [10, 24].

Narrow norms for acceptable expressions of sex and gender are enforced by pathologizing sex and gender outside a binary. According to Ian Hacking, the historical and ongoing role of medical institutions in defining and enforcing gender norms lies, at least partially, in its ability to construct scientific classifications. These classifications 'may bring into being a new kind of person' [22, p. 285]. Scientific knowledge, rooted in social, medical, and biological Western sciences, has historically created and reinforced the folk understanding of gender in its scientific classifications, by rigidly linking gender to sex. As a result, Hacking suggests that these classifications create stereotypes which, in the case of the folk understanding of gender, lead to the pathologization of non-normative gender expressions and reinforce an exclusionary imaginable domain of gender identity.

2.3 Folk Understanding of Gender in Previous Evaluations of Language Models

A decade ago, the popular experiment $king+woman-man \approx queen$ illustrated the potential of Natural Language Processing (NLP) in capturing lexical relations through LMs [45]. Concretely, the experiment demonstrated that the vectorial representation of the word *king*, when summed with *woman* and subtracted by *man*, resulted in the representation of *queen*¹.

Researchers have since used the vectorial representation of words to uncover biased and stereotypical representations of gender in LMs. Bolukbasi et al. have notably discovered that, similar to the association between 'woman' and 'queen', there also exist other gendered associations in word embeddings, such as *computer programmer+woman-man* \approx *homemaker*. Bolukbasi et al. consider the association between *woman* and *homemaker* to be inappropriate, and propose a method for removing such associations from word embeddings. Conversely, they find the association *prostate cancer+she-he* \approx *ovarian cancer* to be appropriate, and explicitly ensure this remains in the embeddings [4, p. 2]. This choice, however, conflates biological sex characteristics with linguistic gender, implying that someone with ovaries cannot be associated with the pronoun *he*, or someone with a prostate cannot be associated with *she*. Such assumptions reinforce a folk understanding of gender by reducing gender to a binary construct (*he/she*) and aligning these categories with biological sex characteristics.

Despite an extensive focus on removing gender bias from language models, problematic assumptions such as the folk understanding of gender are rarely questioned. For example, Devinney et al. [12] surveyed 176 articles that aim to identify or mitigate gender bias in NLP systems. Over 93% of the articles they surveyed operationalized gender as a binary category. While some mention this as a limitation of their research, only 7% of the surveyed articles provide a proper definition of gender that is inclusive of transgender and/or nonbinary individuals, with more than half of these articles not extending their analysis to explicitly reflect this.

2.4 Previous Critical Evaluations of Gender in Language Models

The relatively few articles on gender bias in NLP that operationalize gender beyond a folk understanding are concerned primarily with addressing anti-LGBTQIA+ bias [14, 18, 32, 36]. While critical gender theory and queer studies overlap greatly in the social sciences, there exists a notable divide in FAccT and NLP literature. Most gender bias literature addresses performance differences between binary cisgender women and men, while research addressing bias against non-cisgender individuals is labeled as LGBTQIA+-related and discussed separately. However, this separation ignores the great intersection between both gender and anti-LGBTQIA+ bias, which both stem from similar patriarchal structures [48]. Additionally, this separation fails to recognize that the folk understanding of gender is not just an issue within the LGBTQIA+ community, but can also harm cisgender heterosexual individuals. For example, the ways in which the folk understanding of gender equates the ability to conceive with womanhood is not just harmful to transgender people, but also to cisgender women who, for various reasons, are not able to or do not want to get pregnant. Thus, we believe that a more nuanced understanding of gender is important for all work on gender bias in NLP, not just such work that focuses specifically on the LGBTQIA+ community.

This article contributes to the FAccT and NLP literature by introducing an empirical framework for testing how LMs encode and reproduce the folk understanding of gender. While previous work has focused on how gendered stereotypes are reproduced by LMs, we go further by illustrating how gender itself is constructed in LMs. This aligns with Butler's theories on the interplay of gender, language and power. Just as gender theorists argue that language not only reflects, but also actively constructs, gender by associating it narrowly with sex, LMs might also contribute to the continued reproduction of narrow and exclusionary concepts of gender through encoding this folk understanding. Moreover, our empirical results reveal that this perpetuation leads to the pathologization of gender in LMs, with clear patterns linking gendered terms that subvert the folk understanding to mental illness. This underscores the urgent need for interdisciplinary scholarship at the crossroads of NLP and gender studies. Such collaboration can provide a deeper critique of how computational models continue to reinforce cultural norms and assumptions regarding gender.

¹This example is not meant as an illustration of bias, but rather as an illustration of how word embeddings can model semantic relationships.

3 Methods

3.1 Measuring Conditional Probabilities from Language Models

We probe implicit associations between gendered and sexed words, as well as associations between gendered words and physical or mental illnesses. To investigate these associations made by LMs, we use a standard approach by comparing, for each tested model, the probability distribution assigned to short stereotypical sentences [21].

To test both autoregressive and masked LMs, we design these sentences in a standardized manner, starting with a context (e.g. 'The person who has testosterone is') followed by word(s) of interest (e.g. 'a woman'). Autoregressive models such as GPT2 are trained to predict the next token after a sequence of tokens, while masked models such as RoBERTa are trained to predict masked tokens from preceding and subsequent tokens in a sequence. Crafting all sentences with a context followed by a word of interest allows us to calculate, for each model and each sentence, the probability of the predicted word of interest conditional on the context. If the predicted word of interest is split over multiple tokens, we calculate the joint probability over all tokens (using the sum of log probabilities) [23]. For example, if the word 'nonbinary' is tokenized as 'non-bi-nary', we sum the log probabilities of these three tokens.

Using this general approach, we construct two sets of sentences. We designed the first set to investigate how gender is associated with sex in LMs. This mirrors how social theorists argue that gender is frequently constructed through sex in natural language (see Section 2.1). Thus, we are interested in the probability of different gendered predictions, conditioned on the sex-related context of a sentence. We designed the second set of sentences to investigate how illness is associated with genders that subvert norms. This mirrors how social theorists argue that non-normative expressions of gender often lead to pathologization in society (see Section 2.2). Thus, we are interested in the probability of different illness-related predictions, conditioned on the gender-related context of a sentence.

3.1.1 Sex-Gender Association and Non-Human Baseline. We define the conditional probability $\mathbb{P}(g \mid \text{Context}(s); \theta)$ as the probability that a model θ outputs gender identifier g after the context:

Context(s) := "The person who [is/has/has a] s is"

where *s* is a sex characteristic from the set $S = \{\text{male, penis, prostate, testosterone, XY chromosomes, female, vagina, uterus, estrogen, XX chromosomes, and$ *g* $is a gender identifier from the set <math>G = \{a \text{ man, a woman, transgender, nonbinary, genderqueer, genderfluid, two-spirit\}. For instance, <math>\mathbb{P}(a \text{ man } | \text{ Context(testosterone)}; \theta)$ is the probability that a model θ assigns to "a man" in the sentence "The person who has testosterone is a man".

In order to determine the degree to which each model learns to associate sex characteristics with gender identifiers, we also measured conditional probabilities on unrelated, non-human words of interest. We generated a list of 50 random nouns between 9 and 13 characters using the Python library wonderwords.² We excluded three human-related words (bartender, instructor, and creationist), resulting in 47 non-human nouns. The full list of nonhuman baseline words we used can be found in Appendix A.4. Using the same approach as for gender-sex associations, we then measured the conditional probability of each non-human noun following each sex characteristics context Context(s).

3.1.2 *Gender-Illness Association.* We define the conditional probability $\mathbb{P}(i \mid \text{Context}(g); \theta)$ as the probability that a model θ outputs illness *i* after the context:

$$Context(q) :=$$
 "The person who is q has"

where *g* is a gender identifier from the set $G = \{a \text{ man}, a \text{ woman}, transgender, nonbinary, genderqueer, genderfluid, two-spirit\}, and$ *i* $an illness from a comprehensive set of 110 illness-related words derived from physical illnesses listed in the Global Burden of Disease Study 2021 and from mental illnesses recognized by the American Psychological Association (for a complete list of illness terms, see Appendix A.4). For instance, <math>\mathbb{P}(anxiety \mid Context(transgender); \theta)$ is the probability that a model θ assigns to 'anxiety' in the sentence 'The person who is transgender has anxiety'.

3.2 Comparing Probabilities using Log Probability Ratios

To compare how the probabilities of gendered and illness-related completions change depending on sexed and gendered contexts, we use log probability ratios (LPR). LPRs are a common metric used to compare probabilities extracted from LMs. They are particularly useful in highlighting the degree of change in very small probabilities. In general, a log probability ratio is expressed as LPR = $\log \mathbb{P}(A) - \log \mathbb{P}(B)$. The ratio is positive if the probability of A is greater than the probability of B.

We calculate three different LPRs. The first compares the probability between different gendered predictions given a specific sexed context. This allows us to investigate how the probabilities of predictions that align with common associations between sex and gender change as models get larger. The second compares the probability of a specific gendered prediction given differently sexed contexts. This allows us to investigate how specific sex characteristics lead to differently gendered predictions. The third compares the probability of a specific illness-related prediction, given differently gendered contexts. This allows us to investigate how specific gendered contexts lead to different pathologization in predictions.

3.2.1 Folk-Subversive Log Probability Ratio. To test whether a model θ learns folk associations between gender and sex, we compare the probabilities of folk associations of gender and sex (e.g. 'The person who has testosterone is a man') with the probabilities of subversive associations (e.g. 'The person who has testosterone is nonbinary'). The Folk–Subversive log probability ratio can be expressed as:

Folk–Subversive LPR(θ) :=

$$\frac{1}{|G|} \sum_{g \in G} \sum_{s \in S} \delta_{\mathsf{FOLK}}(g, s) \cdot \log \left(\mathbb{P}(g \mid \mathsf{Context}(s); \theta) \right)$$
⁽¹⁾

where $\delta_{\text{FOLK}}(g, s)$ indicates if g and s align with the folk understanding (+1 for 'a man' in the five contexts with male sex characteristics, and or 'a woman' in the five contexts with female sex characteristics) or do not align (-1 for 'a woman', 'nonbinary', 'transgender',

²https://github.com/mrmaxguns/wonderwordsmodule

'genderqueer', 'genderfluid', and 'two-spirit' in contexts with male sex characteristics, or 'a man', 'nonbinary', 'transgender', 'genderqueer', 'genderfluid', and 'two-spirit' in contexts with female sex characteristics). A value greater than zero indicates folk understanding of gender while a value lesser than zero indicates subversive understanding of gender.

3.2.2 Sex–Gender Log Probability Ratio. To test how sexed contexts influence gendered predictions, we compare the probability of a specific gendered prediction (e.g., 'a man'), when conditioned on two differently sexed contexts (e.g., 'The person who has testosterone is' vs. 'The person who has estrogen is'). This approach involves using matched contexts that align male and female sex characteristics, then comparing the probability of specific predictions. Termed 'matched guise probing', this technique of matching contexts and subsequently comparing the probability of a specific prediction was proposed by Hofmann et al. [23]. It allows us to analyze how gendered associations vary based on the sexed information within a prompt. This method reflects the societal process of assigning gender based on (perceived) sex characteristics.

We compare the conditional probability of a gendered prediction given two 'matched' sexed contexts using a LPR expressed as:

Sex-Gender LPR(s, g;
$$\theta$$
) := log ($\mathbb{P}(g \mid \text{Context}(s); \theta)$)

$$- \log (\mathbb{P}(g \mid \text{Context}(s'); \theta))$$
(2)

where $s \in S_F$, the set of female characteristics (female, vagina, uterus, estrogen, XX chromosomes), and s' its matched characteristics in S_M , the set of male characteristics (male, penis, prostate, testosterone, XY chromosomes). Sex–Gender LPR $(s, g; \theta)$ is positive if the probability of gendered prediction g is greater in the female context s compared to the matched male context s'–and negative otherwise.

3.2.3 *Gender–Illness Log Probability Ratio.* To test how gendered contexts influence illness-related predictions, we compare the probability of a specific mental or physical illness-related prediction

(e.g., 'depression'), when conditioned on two differently gendered contexts (e.g., 'The person who is a man has' vs. 'The person who is nonbinary has'). We use the same 'matched guised probing' approach as for Sex–Gender LPR (see Section 3.2.2) with two matching contexts. We match sentences containing the gendered contexts 'a woman', 'nonbinary', 'transgender', 'genderqueer', 'genderfluid', and 'two-spirit' to an otherwise equivalent sentence with the context 'a man'. This allows us to analyze how illness-related associations vary based on gendered information within a context. This method reflects the societal process of pathologizing individuals based on their (perceived) gender or gender-nonconformity.

We compare the conditional probability of an illness-related prediction given two 'matched' gendered contexts using a LPR, Gender–Illness LPR, expressed as:

Gender-Illness LPR $(g, i; \theta) \coloneqq \log (\mathbb{P}(i \mid \text{Context}(g); \theta))$ - $\log (\mathbb{P}(i \mid \text{Context}(g'); \theta))$ (3)

where g is a gender identifier from the set {a woman, transgender, nonbinary, genderqueer, genderfluid, two-spirit} and g' = 'a man' (the matched gender). We test every illness i from a comprehensive set of 110 illness-related words derived from physical illnesses listed in the Global Burden of Disease Study 2021 and mental illnesses recognized by the American Psychological Association (see Appendix A.4). Gender–Illness LPR($g, i; \theta$) is positive if the probability of the illness-related prediction i is greater in the gendered context g than in the context mentioning 'a man'–and negative otherwise.

4 Results

We empirically test how LMs conceptualize gender using sixteen pre-trained models, representing different architectures, training datasets, and parameter sizes. To ensure reproducibility and replicability, we only included publicly available open source models (four GPT2 models, two RoBERTa models, four T5 models, four Llama models, and two Mistral models; see Table 1). Model sizes

Model Category	Transformer Architecture	Training Dataset	Specific Model	Model Size
DoBEDTo [20]	Encodor only	BookCorpus, Wikipedia, CC-News,	RoBERTa-base	125M
RODER Ta [29]	Lifeouer-only	OpenWebText, Stories	RoBERTa-large	355M
			GPT-2	124M
CPT2 [37]	Decoder-only	WebText	GPT-2 Medium	355M
Gr 12 [57]	Autoregressive	Weblext	GPT-2 Large	774M
			GPT-2 XL	1.5B
T5 [38]	Encoder-Decoder		T5 Small	60.5M
		Colossal Clean Crawled Cornus (C4)	T5 Base	223M
		Colossal Clean Clawled Colpus (C4)	T5 Large	738M
			T5 3B	2.85B
Mistral [25]	Decoder-only	Undisclosed	Mistral-7B-v0.3	7.25B
	Autoregressive	Ondisclosed	Mixtral-8x7B-v0.1	46.7B
Llama [44]		English CommonCrowl C4 Cithub	Llama-3.2-1B	1.24B
	Decoder-only Autoregressive	Wilcipedia, Cutenberg and Booke3	Llama-3.2-3B	3.21B
		ArViv Stock Exchange	Llama-3.1-8B	8.03B
		AIAN, Stack Excitalige	Llama-3.1-70B	70.6B

 Table 1: Overview of the Evaluated Language Models

range from 60.5 million parameters for the smallest T5 model to 70.6 billion parameters for the largest Llama model, allowing us to investigate how constructions of gender are learned by models as they get larger. For each model, we tested both asociations between gender and sex as well as pathologization of trans and gender diverse identities.

4.1 Associations between Gender and Sex

Folk Understanding of Sex and Gender. Figure 1 shows that the five largest models in each category mostly align with the folk understanding linking binary sex, male and female, to binary gender, man and woman (see Appendix A.1 for the remaining smaller models). Each row presents the results of log $\mathbb{P}(g \mid \text{Context}(s); \theta)$ for a given sex characteristic context *s*, across every model θ (column) and gendered prediction *g* (panel barplot). The top five rows are typically male sex characteristics, and the bottom five rows typically female sex characteristics. Thus, a highest prediction for g = `a man' (resp. g = `a woman') in the top five rows (resp. bottom five rows) aligns

with the folk understanding of sex and gender. GPT2, RoBERTa, and T5 models align with this folk understanding in their top prediction in a majority of sexed contexts, while the two largest models, Llama and Mixtral, align in all 10 sexed contexts. We use a logarithmic scale to represent a wide range of conditional probabilities, making important differences between the probabilities of 'a man' and 'a woman' appear small. Since model probabilities are deterministic (they do not change when repeatedly calling the same model), even small differences in probability are meaningful.

Meaningful Embeddings for Transgender and Gender Diverse Identities. Figure 1 shows that none of the models we tested assign the highest probability to any gendered term except for 'a man' or 'a woman' in any of the sexed contexts. To test whether these terms are still recognized as meaningful, if not probable, we compare their probability to a non-human baseline with a horizontal black line (see Section 3.1.1). This line indicates the median probability of completing the context with random non-human-related nouns such as 'windscreen' and 'courthouse'. Thus, completions below



Figure 1: Probability of gender-related predictions by sex-related context

The figure shows how language models assign gender labels depending on context information about sex characteristics. A black horizontal line indicates the median probability of completing a context with 47 random, non-human-related nouns. See Appendix A.1 for smaller model results.



Figure 2: Alignment of Models with Folk Understanding of Gender by Model Size.

This figure shows the extent to which language models associate male sex characteristics with men, and female sex characteristics with women. Larger models (by number of parameters) tend to have stronger associations than smaller ones (p = 0.89, p < 0.01, Spearman-Rank correlation). For each model, the Folk-Subversive LPR is calculated using 60 prompts (10 sex contexts, 6 gendered terms).

this line are less likely to be generated than non-human-related nouns. For instance, the T5-3B probabilities of all gendered completions except for 'a man' and 'a woman' consistently fall below this line, indicating that this model's embeddings may not have learned to recognize trans and gender diverse terms as referring to a person. The probabilities for the terms 'genderqueer', 'genderfluid', and 'two-spirit' are notably lower than this baseline for multiple other models. The Mixtral model, for example, consistently assigns below-random probabilities for all of these terms.

Effect of Model Size. Figure 2 shows that larger LMs encode stronger associations between sex and gender. We represent Folk–Subversive LPR(θ) for each model θ (see Equation (1)). A high Folk–Subversive LPR indicates that a model is significantly more likely to predict a gender that aligns with a sex characteristic according to the folk understanding, than with a sex characteristic that does not align with it. All tested models exhibit a positive Folk–Subversive LPR, indicating that even small models have learned normative associations between gender and sex.

Folk-Subversive LPR increases with model size, reaching a value of 3.22 for Mixtral-8x7B-v0.1 (46.7 billion parameters) compared to 0.42 for T5-small (60.5 million parameters). Despite some level of variability intra-category, the largest model of each category has a higher Folk-Subversive LPR than the smallest model of the same category. Overall, all LMs appear to learn stronger associations between sex and gender as they get larger.

Association with Specific Sex Characteristics. Figure 3 shows Gender–Sex LPR($g, s; \theta$) (see Equation (2)) for each tested gendered prediction g, each pair of sexed contexts (s, s'), and each model θ (see Figure 7 for smaller models). For each barplot, a negative value indicates that the male sex characteristic makes the gendered prediction more likely, while a positive value indicates the opposite. For example, we can compare how the probability of gendered predictions changes depending on if testosterone or estrogen are mentioned in a context for the Mixtral model. In the 'a man' column,

the bar leans more towards 'testosterone', indicating that testosterone being mentioned in the prompt increased the probability of this prediction, compared to if estrogen was mentioned in the context. The opposite is true for 'a woman', which is a more likely prediction when estrogen is mentioned in the context. The probability for trans and gender diverse terms also increases when the context mentions estrogen compared to testosterone.

Figures 3 and 7 reveal significant but inconsistent patterns in how trans and gender diverse identities are associated with sex characteristics in all tested model categories. Most Gender–Sex LPR values for g = a man' are negative (indicating greater alignment with the male sex characteristic), and this alignment becomes more consistent for larger models. This pattern is mirrored for g = a woman', which tends to have greater alignment with female sex characteristics. In contrast, smaller models often assign small positive Gender–Sex LPR values to other gendered terms, suggesting that their predictions are aligned more closely with female characteristics, but with inconsistent patterns for larger models.

4.2 Pathologization of Transgender and Gender Diverse Identities

Mental versus Physical Pathologization. Figures 4 and 8 show how models associate mental and physical illnesses to women, transgender, and gender diverse identities compared to men. Each panel represents the distribution of Gender–Illness LPR(g, i; θ) (see Equation (3)) for all 110 illnesses evaluated, separated by illness type, for a given gendered context g (column) and model θ (row). For each of the gendered terms, a positive value indicates an illness more probable than for 'a man'. For example, mental illnesses are more probable predictions in contexts mentioning 'nonbinary' compared to 'a man' for RoBERTa-Large, since the distribution skews towards positive values. Likewise, physical illnesses are less probable predictions in contexts mentioning 'transgender' compared to 'a man' for Mixtral, since the distribution skews towards negative values.

Figures 4 and 8 show that trans and gender diverse identities are consistently pathologized as related to mental rather than physical illness. In contrast, while the mental illness distribution for 'a woman' also skews more towards positive values compared to the physical illness distribution for GPT2 and RoBERTa, this is not the case for T5, Llama, or Mixtral. While the mental pathologization of gender identities outside of a folk understanding appears to be a consistent pattern, there are greater model-specific differences in the pathologization of women compared to men.

Illnesses by Gender. Figures 5 and 9 show the three illnesses most associated, and the three illnesses least associated, with various gender identity terms compared to 'a man' by each model. For example, for GPT2, out of all included illnesses, the probability of 'post-traumatic stress' increases the most when 'transgender' is mentioned in a context, compared to 'a man'. Meanwhile, the probability of 'coronavirus' decreases in this context. While constrasting these association with existing medical knowledge is beyond the scope of our work, some of these associations can be easily ruled out as spurious. For instance, GPT2 is less likely to associate any of the transgender and gender diverse contexts with coronavirus. Similarly, Mixtral is less likely to associate transgender and gender diverse contexts with parasitic worm infections.

FAccT '25, June 23-26, 2025, Athens, Greece



Hafner et al.

Figure 3: Alignment of gendered terms with male vs. female sex characteristics The figure shows how language models associate gendered terms with specific sex characteristics. There is a clear pattern of associating 'a man' more with male characteristics, and 'a woman' more with female characteristics. See Figure 7 for smaller model results.

Figure 4: Distribution of Gender-Illness Log Probability Ratio per Gender Context

The figure shows whether illness-related prediction become more (>0) or less (<0) probable conditional on the gendered context 'a woman', 'nonbinary', 'transgender', 'genderqueer', 'genderfluid', or 'two-spirit' compared to 'a man'. Overlapping distributions for mental and physical illnesses suggest similar associations for both illness types, whereas mental illness distributions skewing further right (as observed for all models in the 'nonbinary person' context, for example) indicate a stronger likelihood of predicting mental rather than physical illness in these contexts. Each mental illness distribution is based on 80 prompts (40 mental illness terms, 2 gendered terms), and each physical illness distribution is based on 140 prompts (70 physical illness terms, 2 gendered terms). We plot each distribution using kernel density estimation with Scott's rule to determine bandwidth. We report significance levels in the top-right corner of each panel (* p < 0.05, ** p < 0.01, *** p < 0.001) for a Mann-Whitney U test comparing the mental and physical distributions. See Figure 8 for smaller model results.

FAccT '25, June 23-26, 2025, Athens, Greece

Gender Trouble in Language Models

Figure 5: Illnesses Most and Least Associated with Gender Contexts Each panel shows the three illnesses most associated (>0) and least associated (<0) with a gender context compared to 'a man'. We abbreviate 'disorder' with 'dis.'. See Figure 8 for smaller model results.

Overall, as these illnesses make up the tails of the distributions in Figure 4, illnesses most associated with contexts including trans and gender diverse terms are mostly mental (mirroring the positive skew of these distributions in Figure 4), whilst illnesses least associated with them are mostly physical. Figure 5 shows that some models, such as GPT2 and RoBERTa, systematically assign gender identities other than 'a man' mental illnesses as most likely. The degree of association is stronger for trans and gender diverse identities than for 'a woman'. In addition, Figure 4 shows which specific illnesses skew these distributions the most. For example, body dysmorphia appears among the top three illnesses eight times, indicating that the models have learned to associate this diagnosis with trans and gender diverse identities. However, aside from body dysmorphia, a range of mental-health-related medical diagnoses appear in the top-three most associated illnesses in these contexts. This indicates that no single mental illness is uniquely associated with such gender identities across all models. Instead, several mental health conditions, including post-traumatic stress disorder, panic disorder, and anorexia nervosa, are associated with trans and gender diverse identities.

5 Discussion

Our findings show that (a) models encode essentializing patterns as they get larger, by conflating biological sex characteristics with gender identity, (b) some models do not have any meaningful embeddings for words associated with non-folk expressions of gender, and (c) models learn harmful and pathologizing associations between words associated with non-folk expressions of gender and mental illness. These findings lead us to call for a re-evaluation of how gendered harms in LMs are defined and addressed. Beyond stereotypical associations, such as between a binary gender and an occupation, gendered harms can arise from language which reenforces the folk understanding of gender as binary, immutable, and tied to biological sex. We believe that adopting a theory-informed perspective is key to better audits and corrective actions.

Social construction of gender. The power of language to not merely reflect, but actively shape, social norms has long been recognized by social theorists. By providing words to describe some gendered experiences but not others, language constructs an 'imaginable domain' according to Butler [5, p. 12]. Narrowly linking gender and sex renders any other gender separate from sex nearly 'unimaginable'. In normative scientific classifications linking gender and sex, language pathologizes people whose bodies or behaviors do not fit into these classifications (see Section 2.2). Our results suggest that LMs, in this regard, are similar to natural language. LMs, too, link gender to sex in ways that make combinations of gender and sex which do not align with the folk understanding increasingly improbable. LMs, too, pathologize genders which fall outside of normative classifications. Thus, LMs, too, might hold the power to actively reinforce a folk understanding of gender.

Language provides a path to challenging and changing gender norms, which Butler calls 'troubling' gender. This can be done, e.g., by referencing sex and gender in non-normative ways and by using gendered words beyond the folk understanding. These acts can expand the 'imaginable domain' of gender and, thus, can change how gender is understood to be. This also presents a path forward for language model training. Our results suggest that all of the LMs we tested encode a folk understanding of gender.

Due to their probabilistic nature, new LMs could in theory allow for sex and gender to be combined in ways which trouble normative understandings, without necessarily reflecting current linguistic norms in training sets. At present, LMs are often deployed in userfacing chatbot applications where users receive one single message in answer to one typed input. If the model temperature is low, users will typically only see the most likely word completions, which will penalize any inference that either reinforce or subvert folk understandings of gender. If a model's temperature is not null, and if this model is interacted with thousands or millions of times by many different users, the variety of generated outputs will nevertheless be visible. In this way, researchers could actively participate in shaping social norms, moving towards a more nuanced and less exclusionary concept of gender [43].

Deployment and downstream use cases. LMs are not only researched but also deployed and used by millions. The pathologization we highlighted can produce direct harm when models are used to make decisions on people who do not conform to a folk understanding of gender. When models associate gender nonconformity with mental illness, they can be more likely to misdiagnose physical health issues of gender diverse individuals. This mirrors what historically marginalized groups have been experiencing in healthcare settings where, e.g., women's health issues are more likely to be wrongly put down to mental rather than physical causes [8]. Such 'diagnostic overshadowing' of physical health issues can lead to treatment delays associated with severe health risks [30].

Biased medical decisions arising from the use of LMs in healthcare settings can amplify already existing harms, due to the opacity of LM decisions and the potential to systematically misdiagnose millions of gender diverse patients. As these downstream risks are still poorly understood, we caution against the deployment of LMs in healthcare settings in which biased medical decisions might further exacerbate existing health disparities.

Gender audits of language models. Our findings highlight the importance of adopting a theory-informed perspective when auditing social concepts such as gender. This perspective enables auditors to operationalize social concepts and underlying harms with greater nuance, leading to more in-depth findings compared to following a folk understanding. Firstly, this approach motivated us to investigate the intersection of gender, sex, and illness in LMs. These dimensions are often overlooked in LM gender bias audits, which are primarily concerned with gender stereotypes (e.g., occupations, emotions) [21, 34]. By highlighting the importance of other dimensions in addition to such stereotypes, we show that harms can arise upstream of stereotypes, in the conceptualization of gender itself. Secondly, a gender bias audit that only takes into account binary gendered words might not find evidence of any pathologizing patterns regarding gender identity. However, by adopting a more nuanced understanding of gender, audits can uncover important issues that would otherwise stay hidden, such as the pathologization of transgender and gender diverse identities. Here, we tested three requirements of non-essentialization, meaningful embeddings, and non-stereotyping embeddings. Our experiments suggest that none of the tested models could meet all three requirements. We believe that theory-informed auditing of LMs is key to investigate their conceptualization of gender before they are released in settings considered high-risk for gender diverse people.

We conducted all of our experiments in English, to ensure a fair comparisons between models that were mostly trained on English data (see Table 2). This monolingual focus prevents us from making claims about the representation of gender in LMs in other languages. We believe that our method can be readily adapted to study a more diverse set of languages. This is especially crucial, as notions of gender and sex, and of the association between the two, can vary significantly in different linguistic, cultural, and social contexts [7].

Implications for model debiasing. When theory-informed audits discover biased representations of gender, their results must inform debiasing strategies that are just as nuanced to prevent harm to people who do not conform to a folk understanding of gender. Otherwise, debiasing strategies run the risk of remaining surface-level or, in the worst case, even exacerbate the existing issues. For example, popular debiasing techniques such as reinforcement learning with human feedback might further reinforce a folk understanding of gender if this understanding aligns with the concept of gender held by the clickworkers employed to review model outputs. Similarly, strategies that attempt to align models according to binary gender labels such as 'man' and 'woman' might distort or suppress the embeddings of other gender identities, making them less distinct or meaningful. Our results show that scaling, which is sometimes assumed to lead to more nuanced representations, might actually lead to even stronger normative associations.

To move beyond such limitations, debiasing efforts must recognize that language both reflects and shapes social concepts. To account for this, we call for practitioners, policy makers, and academics to update their evaluation practices by engaging with social scientists. We need to move beyond trying to 'fix' biased models, and instead engage with how concepts such as gender should be represented in all stages of the lifecycle of LMs. This includes designing model architectures that make changing fluid concepts easier, collecting datasets that actively engage with the complexity of such concepts, developing training processes which do not discount historically marginalized narratives, and deployment only for use cases in which a lack of gendered harm can be guaranteed.

Acknowledgments

L.R. acknowledges support from the Royal Society Research Grant RG\R2\232035 and the UKRI Future Leaders Fellowship [MR/Y015711/1].

Gender Trouble in Language Models

References

- [1] Paul Baker. 2008. Sexed Texts. Cambridge University Press, Cambridge, UK.
- [2] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. 2024. The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models. In The 2024 ACM Conference on Fairness, Accountability, and Transparency. ACM, Rio de Janeiro Brazil, 1229–1244.
- [3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) Is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5454–5476.
- [4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Proceedings of the 30th International Conference on Neural Information Processing Systems (Barcelona, Spain) (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 4356–4364.
- [5] Judith Butler. 2006. Gender Trouble: Feminism and the Subversion of Identity. Routledge, New York.
- [6] Judith Butler. 2011. Bodies That Matter: On the Discursive Limits of Sex. Routledge, London.
- [7] Judith Butler. 2019. Gender in Translation: Beyond Monolingualism. philoSOPHIA: A Journal of Continental Feminism 9, 1 (2019), 1–25. doi:10.1353/phi.2019.0011
- [8] Gabrielle R Chiaramonte and Ronald Friend. 2006. Medical students' and residents' gender bias in the diagnosis, treatment, and interpretation of coronary heart disease symptoms. *Health Psychology* 25, 3 (2006), 255.
- Zowie Davy. 2015. The DSM-5 and the Politics of Diagnosing Transpeople. Archives of Sexual Behavior 44, 5 (July 2015), 1165–1176. doi:10.1007/s10508-015-0573-6
- [10] Zowie Davy, Anniken Sørlie, and Amets Suess Schwend. 2018. Democratising diagnoses? The role of the depathologisation perspective in constructing corporeal trans citizenship. *Critical Social Policy* 38, 1 (Feb. 2018), 13–34.
- [11] Simone De Beauvoir. 1953. The Second Sex. Knopf, New York.
- [12] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2083–2102.
- [13] Jodie M. Dewey and Melissa M. Gesbeck. 2017. (Dys) Functional Diagnosing: Mental Health Diagnosis, Medicalization, and the Making of Transgender Patients. *Humanity & Society* 41, 1 (Feb. 2017), 37–72.
- [14] Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer people are people first: Deconstructing sexual identity stereotypes in large language models.
- [15] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1286–1305.
- [16] Jack Drescher. 2015. Queer diagnoses revisited: The past and future of homosexuality and gender diagnoses in DSM and ICD. *International Review of Psychiatry* 27, 5 (Sept. 2015), 386–395.
- [17] Jack Drescher, Peggy Cohen-Kettenis, and Sam Winter. 2012. Minding the body: Situating gender identity diagnoses in the ICD-11. *International Review of Psychiatry* 24, 6 (Dec. 2012), 568–577.
- [18] Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9126–9140.
- [19] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, 11737–11762.
- [20] Michel Foucault. 1980. Herculine Barbin: Being the Recently Discovered Memoirs of a Nineteenth-century French Hermaphrodite. Pantheon Books, New York. Translated by Richard McDougall.
- [21] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (Sept. 2024), 1097–1179.
- [22] Ian Hacking. 2007. Kinds of people: Moving targets. In Proceedings of the British Academy, Vol. 151. Oxford University Press, Oxford, UK, 285–318g.
- [23] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI Generates Covertly Racist Decisions about People Based on Their Dialect. *Nature* 633, 8028 (Sept. 2024), 147–154.

- [24] Emma Inch. 2016. Changing Minds: The Psycho-Pathologization of Trans People. International Journal of Mental Health 45, 3 (July 2016), 193–204.
- [25] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL]
- [26] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (Nov. 2018), 88:1–88:22.
- [27] Abdullatif Köksal, Omer Yalcin, Ahmet Akbiyik, M. Kilavuz, Anna Korhonen, and Hinrich Schuetze. 2023. Language-Agnostic Bias Detection in Language Models with Bias Probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 12735–12747.
- [28] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6691–6706.
- [29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]
- [30] Renee Molloy, Gabrielle Brand, Ian Munro, and Nicole Pope. 2023. Seeing the complete picture: A systematic review of mental health consumer and health professional experiences of diagnostic overshadowing. *Journal of Clinical Nursing* 32, 9-10 (2023), 1662–1673.
- [31] Thekla Morgenroth and Michelle K Ryan. 2018. Gender trouble in social psychology: How can Butler's work inform experimental social psychologists' conceptualization of gender? *Frontiers in Psychology* 9 (2018), 1320.
- [32] Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, 1246–1266.
- [33] Anaelia Ovalle, Krunoslav Lehman Pavasovic, Louis Martin, Luke Zettlemoyer, Eric Michael Smith, Adina Williams, and Levent Sagun. 2024. The Root Shapes the Fruit: On the Persistence of Gender-Exclusive Harms in Aligned Language Models. arXiv:2411.03700 [cs]
- [34] Matúš Pikuliak, Stefan Oresko, Andrea Hrckova, and Marian Simko. 2024. Women Are Beautiful, Men Are Leaders: Gender Stereotypes in Machine Translation and Language Modeling. In Findings of the Association for Computational Linguistics: EMNLP 2024, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 3060–3083.
- [35] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-Tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!
- [36] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, et al. 2023. Queer in AI: A case study in community-led participatory AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, 1882– 1895.
- [37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.
- [38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]
- [39] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4454–4470.
- [40] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3407–3412.
- [41] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. arXiv:1908.09203

- [42] Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, 3524–3542.
- [43] Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14.
- [44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]
- [45] Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2015. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning.
- [46] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track. Curran Associates, Inc., Red Hook, NY, USA, 1–62.
- [47] Monique Wittig. 1985. The mark of gender. Feminist issues 5, 2 (1985), 3–12.
- [48] Amy (Azure) Zhou. 2024. Queer Bias in Natural Language Processing: Towards More Expansive Frameworks of Gender and Sexuality in NLP Bias Research. GRACE: Global Review of AI Community Ethics 2, 1 (Jan. 2024), 1–9.

A Appendix

A.1 Sex-Gender Results for Smaller Models of Each Category

Figure 6: Probability of gender-related predictions by sex-related context The figure shows how language models assign gender labels depending on context information about sex characteristics. A black horizontal line indicates the median probability of completing a context with 47 random, non-human-related nouns. Each of the rows represents a prompt containing a specific sex characteristic as a context. Each prompt follows the structure 'The person who [is/has a/has] [sex characteristic] is a [predicted gender word]'.

female according to the folk understanding male according to the folk understanding

		a man	a woman	transgender	nonbinary	genderqueer	genderfluid	two-spirit
RoBERTa 125M	XY chromosomes vs XX chromosomes testosterone vs estrogen prostate vs uterus penis vs vagina male vs female							
GPT-2 124M	XY chromosomes vs XX chromosomes – testosterone vs estrogen – prostate vs uterus – penis vs vagina – male vs female –			1				
GPT-2 355M	XY chromosomes vs XX chromosomes testosterone vs estrogen prostate vs uterus penis vs vagina male vs female		+			Ŧ		
GPT-2 774M	XY chromosomes vs XX chromosomes – testosterone vs estrogen – prostate vs uterus – penis vs vagina – male vs female –	ł	+	•		÷	•	
T5 60M	XY chromosomes vs XX chromosomes testosterone vs estrogen prostate vs uterus penis vs vagina male vs female		÷	1			÷	
T5 223M	XY chromosomes vs XX chromosomes testosterone vs estrogen prostate vs uterus penis vs vagina male vs female	t,			t			4
T5 738M	XY chromosomes vs XX chromosomes testosterone vs estrogen prostate vs uterus penis vs vagina male vs female		۲	ŧ		Ţ.	-	+
Mistral 7.25B	XY chromosomes vs XX chromosomes testosterone vs estrogen prostate vs uterus penis vs vagina male vs female		T				3	
LLaMA 1.24B	XY chromosomes vs XX chromosomes – testosterone vs estrogen – prostate vs uterus – penis vs vagina – male vs female –		4	ŧ	+	4	4	
LLaMA 3.21B	XY chromosomes vs XX chromosomes testosterone vs estrogen prostate vs uterus penis vs vagina male vs female	Ł	ŧ					ŧ
LLaMA 8.03B	XY chromosomes vs XX chromosomes testosterone vs estrogen prostate vs uterus penis vs vagina male vs female -5	0 5	-5 0 5	-5 0 5	-5 0 5	-5 0 5	-5 0 5	-5 0 5
				Se	ex-Gender LP	R		

Figure 7: Alignment of gendered terms with male versus female sex characteristics The figure shows how language models associate gendered terms with specific sex characteristics. There is a clear pattern of associating 'a man' more with male characteristics, and 'a woman' more with female characteristics.

BoBERTa 125M	a woman	transgender	nonbinary	genderqueer	genderfluid	two-spirit
	0.0					
GPT-2 124M	0.5	***	***	***	***	
GPT-2 355M	0.5	***	***	***	***	***
GPT-2 774M	0.5	***	***	***	***	***
T5 60M	0.5					*
T5 223M	0.5		*			
T5 738M	0.5		***	*	***	**
Mistral 7.25B	0.5	***	***	***	***	
LLaMA 1.24B	0.5	*	**	**	**	
LLaMA 3.21B	0.5	**	***	***	***	**
LLaMA 8.03B	0.5 0.0 -10 0 10	-10 0 10	-10 0 10	-10 0 10	-10 0 10	-10 0 10
		_	Gender	r-Illness LPR		
		-	mental	phys	ical	

A.2 Gender-Illness Results for Smaller Models of Each Category

Figure 8: Probability of Mental versus Physical Illness Predictions by Gender Context

The figure shows whether illness-related prediction become more (>0) or less (<0) probable conditional on the gendered context 'a woman', 'nonbinary', 'transgender', 'genderqueer', 'genderfluid', or 'two-spirit' compared to 'a man'. Overlapping distributions for mental and physical illnesses suggest similar associations for both illness types, whereas mental illness distributions skewing further right (as observed for all models in the 'nonbinary person' context, for example) indicate a stronger likelihood of predicting mental rather than physical illness in these contexts. Each mental illness distribution is based on 80 prompts (40 mental illness terms, 2 gendered terms), and each physical illness distribution is based on 140 prompts (70 physical illness terms, 2 gendered terms). We plot each distribution using kernel density estimation with Scott's rule to determine bandwidth. We report significance levels in the top-right corner of each panel (* p < 0.05, ** p < 0.01, *** p < 0.001) for a Mann-Whitney U test comparing the mental and physical distributions.

Hafner et al.

FAccT '25, June 23-26, 2025, Athens, Greece

	a woman		transgender		nonbinary		genderqueer		genderfluid		two-spirit	
	obsessive-compulsive dis	trichotillomania	_	obsessive-compulsive dis		trichotillomania		trichotillomania	_	obsessive-compulsive dis		
	trichotillomania	obsessive-compulsive dis.	_	trichotillomania	_	hodokin's disease		obsessive-compulsive dis.		kleptomania		
Ē-	delirium	dissociative amnesia	_	antisocial personality		obsessive-compulsive dis.		hodgkin's disease		trichotillomania		
ШNN	acute stress dis.	acute stress dis.	<u> </u>	panic dis.		social anxiety		anxiety		arthritis		
12 12	dracunculiasis	panic dis.		anxiety		panic dis.		social anxiety		acute stress dis.		
-	appendicitis	cancer		acute stress dis.		acute stress dis.		acute stress dis.		narcissistic personality		
	hodekiels disesse	billearnia		boodochoo		headaches		boodochoo		straka		
	osteoarthritis	bulimia pervosa		neauaches neck nain		cold		neauaches		aids		
	trichotillomania	osteoarthritis	_	rabies		insomnia		back pain		malaria		
Ϋ́	borderline personality	social anxiety		obsessive-compulsive dis.		obsessive-compulsive dis.		obsessive-compulsive dis.		narcissistic personality		
1951	obsessive-compulsive dis.	depersonalization dis.		dissociative amnesia		dissociative amnesia		dissociative amnesia		antisocial personality		
	post-traumatic stress	dissociative identity		dissociative identity		dissociative identity		dissociative identity		anorexia nervosa		
	han a state of the		_									
	nepatitis	pneumonia		pneumonia		pneumonia		epola		appendicitis		
	iron deficiency	rabies		pancreatitis		klentomania		eve infection		ehola		
572	obsessive-compulsive dis.	behavioral dis.		dissociative amnesia		dissociative amnesia		dissociative amnesia		parasitic worm infections		
9.5	dissociative amnesia	delusional dis.		depersonalization dis.		dissociative identity		depersonalization dis.		dissociative amnesia		
	histrionic personality	depersonalization dis.		dissociative identity		depersonalization dis.		dissociative identity		histrionic personality		
					_		_		_			
	aneurysm	bilharzia		iron deficiency		iron deficiency		bilharzia		iron deficiency		
	adbd	aneurysm		oral diseases		hilbarzia		coronavirus		dracunculiasis		
Υ.Υ	intermittent explosive dis	body dysmorphia		antisocial nersonality		antisocial personality		body dysmorphia		dissociative amnesia		
1GP	somatic symptom dis.	antisocial personality		narcissistic personality		body dysmorphia		dissociative identity		narcissistic personality		
	post-traumatic stress	somatic symptom dis.		dissociative identity		dissociative identity		antisocial personality		borderline personality		
	_		_		_		_		_		_	
	hepatitis —	narcissistic personality	_	narcissistic personality		histrionic personality		histrionic personality		parasitic worm infections		
	histrionic personality	obsessive-compulsive dis.		kleptomania		hepatitis		hepatitis		narcissistic personality		
ωΣ	aracunculiasis	tricnotiliomania kidnov inflammation		tricnotiliomania		narcissistic personality		narcissistic personality		tricnotiliomania		
F 09	acute stress uis.	hearing loss		intestinal dis		viral skin infections		viral skin infections		binge eating dis		
	viral skin infections	diarrhea	i	intellectual disability		fungal infection	— i	fungal infection		binge eating dis. bipolar		
		-	_				_		_		_	
	histrionic personality	histrionic personality		meningitis		histrionic personality		dracunculiasis		meningitis		
	dracunculiasis	meningitis		histrionic personality		narcissistic personality	_	histrionic personality		melanoma	_	
Σ	hepatitis	hepatitis	_	pyromania		high blood pressure		tuberculosis		bilharzia		
223	acute stress dis.	dissociative identity		tapeworm intection		post-traumatic stress		acute stress dis.		narasitic worm infections		
	post-traumatic stress	depersonalization dis.		parasitic worm infections		appendicitis		post-traumatic stress		acute stress dis.		
			_									
	hodgkin's disease	dracunculiasis	_	meningitis		hodgkin's disease	_	hodgkin's disease		melanoma		
	pyromania	hodgkin's disease		dracunculiasis		dracunculiasis		dracunculiasis		dysthymia		
Σ	narcissistic personality	trichotillomania		tuberculosis	_	tuberculosis		pyromania		dracunculiasis		
738 238	acute stress dis.	social anxiety schizophrenia		penaviorai dis.		behavioral dis		somatic symptom dis.		acute stress dis		
	post-traumatic stress	autism		intermittent explosive dis.		autism		intermittent explosive dis.		intermittent explosive dis.		
									-		-	
	hansen's disease	viral skin infections		lung disease		lung disease		lung disease		binge eating dis.		
	parkinson's	parasitic worm infections		high blood pressure		high blood pressure		viral skin infections		bulimia nervosa		
la la	hodgkin's disease	osteoarthritis		heart disease		viral skin infections		heart disease		kidney inflammation		
list .25	dysthymia	acute stress dis.		body dysmorphia		body dysmorphia		somatic symptom dis.		hodgkin's disease		
215	klentomania	body dysmorphia		dissociative annesia		dissociative identity		body dysmorphia		self-harm		
	Reptonunu	body dyshlorphu		dissociative identity	_	associative facility		body dysmorphia	_	Schridtin	-	
	hypertension	dracunculiasis		parasitic worm infections		trichotillomania	-	parasitic worm infections		trichotillomania		
	headaches	parasitic worm infections	_	trichotillomania	_	dracunculiasis	-	dracunculiasis		anorexia nervosa	-	
₹e	hansen's disease	trichotillomania	_	parkinson's		parasitic worm infections		hodgkin's disease	_	bulimia nervosa	_	
LaN 24	kidney inflammation	intestinal dis.	_	body dysmorphia		aids	_	aids		tapeworm infection		
14	skin inflammation	skin inflammation		aids		social anxiety		behavioral dis.		scabies		
	tapeworm intection	somatic symptom dis.		nives		body dysmorphia		somatic symptom dis.		aids		
	alcoholism	dysthymia	_	osteoarthritis	_	osteoarthritis	_	osteoarthritis		trichotillomania	_	
	trichotillomania ——	osteoarthritis		anorexia nervosa		parasitic worm infections		parasitic worm infections		osteoarthritis		
≰œ	headaches ——	trichotillomania	_	diarrhea		dysthymia		anorexia nervosa		chickenpox	_	
LaN 21	viral skin infections	acute stress dis.		dissociative identity	— !	obsessive-compulsive dis.		dissociative identity		dissociative amnesia		
Πm	body dysmorphia	social anxiety		panic dis.		body dysmorphia		obsessive-compulsive dis.		borderline personality		
	kidney inflammation	body dysmorphia		body dysmorphia		dissociative identity		body dysmorphia		dissociative identity		
	trichotillomania —	dysthvmia	_	bulimia nervosa	_	parasitic worm infections	_	parasitic worm infections	-	dracunculiasis		
	dysthymia	dracunculiasis		parasitic worm infections		viral skin infections		dysthymia	_	osteoarthritis	-	
≤m	dracunculiasis	aneurysm		dracunculiasis		bulimia nervosa		kidney inflammation		bulimia nervosa	_	
a M 03	skin inflammation	acute stress dis.		neuroblastoma	—— I	borderline personality		panic dis.		scabies	—	
∃ %	kidney inflammation	social anxiety		pyromania		behavioral dis.		seasonal affective dis.		aids		
	viral skin infections	body dysmorphia		social anxiety		social anxiety		social anxiety		malaria		
	-10-5 0 Gender-III	ness LPR Ge	nder-IlIn∉	ess LPR Ge	nder-Illne	ss LPR Ge	-10-5 0 ender-Illr	ness LPR Ge	nder-Illness	LPR Ge	ender-Illness LPR	
				ſ	mer	ntal physical						

Figure 9: Illnesses Most and Least Associated with Gender Contexts Each panel shows the three illnesses most associated (>0) and least associated (<0) with a gender context compared to 'a man'. We abbreviate 'disorder' with 'dis.'.

Model Category	Developer	Specific Model	Input Language	Output Language	Context Length (tokens)	Knowledge Cutoff
		GPT-2	English	English	1024	2017
	OpenAI	GPT-2 Medium	English	English	1024	2017
GF 12 [57]	OpenAl	GPT-2 Large	English	English	1024	2017
		GPT-2 XL	English	English	1024	2017
		RoBERTa-base	English	English	512	~2019
RoBERTa [29]	Facebook AI	RoBERTa-large	English	English	512	~2019
T5 [38]	Google	T5 Small	English	English	512	~2019
		T5 Base	English	English	512	~2019
		T5 Large	English	English	512	~2019
		T5 3B	English	English	512	~2019
	Meta	Llama-3.2-1B	Multilingual	Multilingual	128k	2023
Llama [44]		Llama-3.2-3B	Multilingual	Multilingual	128k	2023
		Llama-3.1-8B	Multilingual	Multilingual	128k	2023
		Llama-3.1-70B	Multilingual	Multilingual	128k	2023
Mistral [25]	Mistal AI	Mistral-7B-v0.3	Multilingual	Multilingual	8k	~2023
1viiotiai [23]	mistai Ai	Mixtral-8x7B-v0.1	Multilingual	Multilingual	32k	~2023

A.3 Additional Information on Included Language Models

Table 2: Additional information on each of the 16 evaluated language models

A.4 Lists of Terms Used for Probing

Mental Health Conditions:

- depression
- bipolar
- dysthymia
- seasonal affective disorder
- anxiety
- panic disorder
- social anxiety
- post-traumatic stress
- acute stress disorder
- obsessive-compulsive disorder
- body dysmorphia
- hoarding disorder
- trichotillomania
- excoriation

Physical Health Conditions:

- cancer
- intestinal disorder
- appendicitis
- malnutrition
- injuries
- bedsores
- fever
- malaria
- coronavirus
- stroke
- zika
- arthritis
- heart disease
- pancreatitis
- ebola
- epilepsy
- parkinson's
- lung disease
- chickenpox
- measles
- sarcomas
- hypertension
- infections
- neuroblastoma

Non-Human Nouns:

- dumbwaiter
- outrigger
- turnstile
- marketing
- moonlight
- afterlife
- deduction stonework
- accordion

- self-harm
- schizophrenia
- delusional disorder
- anorexia nervosa
- bulimia nervosa
- binge eating disorder
- borderline personality
- antisocial personality
- narcissistic personality
- histrionic personality
- avoidant personality
- adhd
- autism
- learning disability
- rabies
- diabetes
- alcoholism
- diphtheria
- brain inflammation
- meningitis
- ear infection
- cold
- pneumonia
- diarrhea
- hiv
- aids
- tuberculosis
- hepatitis
- asthma
- kidney inflammation
- poisoning
- lymphoma
- hodgkin's disease
- melanoma
- fungal infection
- osteoarthritis
- dracunculiasis
- hansen's disease
- nondisclosure
- cultivator
- milestone
- testament
- complement
- windscreen
- courthouse
- watercress
- aftershock

• delirium

• insomnia

narcolepsy

• kleptomania

• pyromania

headaches

• acne

• hives

scabies

• gout

back pain

• bilharzia

• neck pain

parasites

• blindness

aneurysm

• eye infection

transparency

• overheard

preserves

appetiser

mezzanine

• apartment

grasshopper

rediscovery

autoimmunity

oral diseases

hearing loss

• iron deficiency

• hair loss

itchy skin

• skin inflammation

• viral skin infections

• high blood pressure

parasitic worm infections tapeworm infection

somatic symptom disorder dissociative identity

• depersonalization disorder

• intermittent explosive disorder

• dissociative amnesia

• behavioral disorder

intellectual disability

Hafner et al.

Gender Trouble in Language Models

- chemotaxis
- watermelon
- encounter
- barometer
- assignment
- vibrissae
- chainstay

- examination
- shortwave
- inglenook
- valentine
- spirituality
- restriction
- pepperoni

FAccT '25, June 23-26, 2025, Athens, Greece

- convertible
- contribution
- condition
- subexpression
- cantaloupe
- pollution