

We Don't Talk About That: Case Studies on Intersectional Analysis of Social Bias in Large Language Models

Hannah Devinney
Dept. Computing Science
Umeå Centre for Gender Studies
Umeå University
hannahd@cs.umu.se
they/them

Jenny Björklund
Centre for Gender Research
Uppsala University
jenny.bjorklund@gender.uu.se
she/her

Henrik Björklund
Dept. Computing Science
Umeå University
henrikb@cs.umu.se
he/him

Abstract

Despite concerns that Large Language Models (LLMs) are vectors for reproducing and amplifying social biases such as sexism, transphobia, islamophobia, and racism, there is a lack of work qualitatively analyzing *how* such patterns of bias are generated by LLMs. We use mixed-methods approaches and apply a feminist, intersectional lens to the problem across two language domains, Swedish and English, by generating narrative texts using LLMs. We find that hegemonic norms are consistently reproduced; dominant identities are often treated as ‘default’; and discussion of identity itself may be considered ‘inappropriate’ by the safety features applied to some LLMs. Due to the differing behaviors of models, depending both on their design and the language they are trained on, we observe that strategies of identifying “bias” must be adapted to individual models and their socio-cultural contexts.

Content warning: This research concerns the identification of harms, including stereotyping, denigration, and erasure of minoritized groups. Examples, including transphobic and racist content, are included and discussed.

1 Introduction

The use of Large Language Models (LLMs) in a wide variety of Natural Language Processing (NLP) tasks and tools, from chatbots to summarization to coreference resolution, is increasing as such models become more widely available both freely and commercially. In such a context, the presence and potential amplification of social biases is of particular concern.

We evaluate the presence and implications of representational harms in the output of LLMs. This is demonstrated for both English (using Llama) and Swedish (using GPT-SW3). After using the LLMs to generate stories, we analyze the resulting corpora for representational harms like stereotyping and denigration. We use the EQUITBL method for

distant (Devinney et al., 2020b) and close readings (Devinney et al., 2020a).

Our main research question asks, *to what extent do LLMs reflect power asymmetries, including intersectional power asymmetries, in the texts they generate?* In particular, we investigate stereotypes, hegemonic norms, erasure of identity in narratives generated by LLMs. We demonstrate how different methods may be necessary to identify and understand biases across models and socio-linguistic settings, due to divergent behaviours.

1.1 Large Language Models

Large Language Models are pretrained on massive amounts of unstructured, unlabeled text data. Transformer-based LLMs are capable of generating text based on patterns discovered within this training data, and can be applied to any tasks which can be rephrased as text generation. We select two open-source LLMs, GPT-SW3 and Llama 2, as case studies to explore different methods for identifying representational harms. This selection allows us to investigate two different linguistic contexts (Swedish and English, respectively), and allows others to reproduce our results.

GPT-SW3. GPT-SW3¹ is a collection of pre-trained LLMs for North Germanic languages, including Swedish, released in 2023 by AI Sweden. From late 2023 it has been made freely available (Ekgren et al., 2023). It has been trained on the Nordic Pile, which contains 1.2 terabytes of text data in Danish, Icelandic, Norwegian, Swedish, and also English: by volume, most of this data is in English and Swedish.

Llama 2. Llama 2² is a collection of open-source, pre-trained LLMs for English released by Meta in 2023. Llama 2 is pretrained for 1.7 million GPU

¹<https://www.ai.se/en/project/gpt-sw3/>

²<https://llama.meta.com/llama2/>

hours with 2 trillion Byte-Pair Encoded tokens from “publicly available sources” with the “most factual” sources upsampled (Touvron et al., 2023). It is then further fine-tuned using Reinforcement Learning with Human Feedback (RLHF), which rewards the model for producing texts preferred by humans. In addition to RLHF, safety is “distilled” into the model by retraining on texts that were generated with prompts focusing on safety.

2 Related Work

2.1 LLMs and Bias

As we might expect based on undesirable system behaviors from other language models that have ‘inherited’ social and historical biases, there are significant concerns about bias in LLMs (see, for example: Felkner et al. (2023); Cheng et al. (2023); Esiobu et al. (2023)). Large language models have been shown to perform worse for gender-neutral pronouns in Danish, English, and Swedish than for gendered pronouns, measured both with respect to intrinsic measures such as perplexity and on several downstream tasks (Brandl et al., 2022). This may in part be due to the ways that tokenization is generally performed in LLMs, and the scarcity of such pronouns in the training data, as shown for English neopronouns (Ovalle et al., 2023).

There are also concerns about LLMs (re)producing other representational harms such as stereotyping or denigration (see, e.g., Felkner et al. (2023); Deas et al. (2023); Venkit et al. (2023)).

2.2 Identifying Bias in Text Corpora

Concannon et al. (2018) use unsupervised topic modeling for feminist analysis of text data, but we prefer a semi-supervised approach to allow us to guide our analysis with respect to the specific groups and power asymmetries we investigate. We therefore follow the EQUITBL method described by Devinney et al. (2020b) and use semi-supervised topic modeling to discover associations between identity groups and particular terms, as well as to identify documents of interest for close-reading to understand the exact nature of such associations.

3 Bias Statement

We consider the overarching concept of ‘bias’ as the concern for how societal power structures manifest in language technologies. With respect to machine-generated narratives, we locate most of

the ‘bias’ we are concerned with investigating under the umbrella of representational harms, particularly stereotyping and erasure.

However, the LLMs we examine do not always return a narrative text when we prompt them to generate one. Thus, we identify several specific system behaviours which we consider distinct harms:

1. *Systematic refusal to answer innocuous prompts.* This behavior constructs some identities, and the concept of “identity”, as risky.
2. *Invalidation of identities.* A subset of (1), when particular terms referring to identities are described as inappropriate, incorrect, or “unimaginable.” This behaviour implies to users who may identify with these terms that they themselves are not welcome in society.³

We investigate identity categories of *gender*, *transness* or trans identity, *race* or ethnicity, and *religion*; as well as (binary) intersectional identities across these categories. All of these social categories constitute and are constituted by the underlying power relations of society, and are inevitably tangled together (Butler, 1999; Crenshaw, 1991; Phoenix, 2006). They are multidimensional, socially constructed, and should not be treated as fixed attributes of individuals (Hanna et al., 2020).

The groups selected in each category are intended to capture power dynamics which have similar asymmetries across both socio-linguistic contexts, and we reduce all dynamics into specific relationships which we think are also comparable across our contexts: anti-trans, anti-Black, and anti-muslim language and attitudes are concerning, and current, in both Swedish and English.

We represent gender with three categories (feminine, masculine, and nonbinary). We consider transness as the misalignment (transgender) or alignment (cisgender) between one’s gender identity and the gender-sex one was assigned at birth. We select binary power relations for race (Black and white) and religion (Muslim and Christian), except in the case study of how race is constructed in the Swedish LLM, where we consider three categories: black, white, and arab; see Section 5.4.

Strictly speaking, ‘arab’ refers to ethnicity rather than race, but race overlaps and intersects with other power asymmetries, such as ethnicity, religion, nationality, and class. Moreover, in the

³Note, however, that claiming an identity is a fairly popular jail-breaking strategy.

Swedish context ethnicity is often used as a proxy for race. Sweden has distanced itself from its race-biological past, which culminated in the establishment of the Swedish Institute for Race Biology at Uppsala University in 1922, through a kind of official color-blindness and refusal to talk about race. Instead, ethnicity is used to discuss differences between groups, but as many researchers have shown, race still figures in these constructions of ethnicity (see, for instance, [Hübinette and Lundström \(2014\)](#)). In this case study, we have chosen to use the category ‘arab’ rather than ‘Muslim’ in an effort to separate the conflated race-ethnicity category from religion.

4 Experimental Setup

Texts generated by LLMs are presented as the object of study in this paper. To obtain them, we develop a broad set of prompts with the intention of capturing a wide array of stories and circumstances. We vary three elements of the input: the *system prompt*, the *template*, and the *noun phrase*. We then generate texts based on the settings described later in this section. All text generation is performed with the *Llama-2-70b-chat-hf* (70 billion parameters, fine-tuned for performance as a chatbot with RLHF additionally applied) and *GPT-SW3-20b-instruct* (20 billion parameters, fine-tuned for performance as a chatbot using instruction tuning) models. Additional materials are also generated for an exploratory investigation of race in texts produced by GPT-SW3, described in Section 5.4.

4.1 System Prompts

The system prompt describes the ‘persona’ of the LLM, and is prepended in front of every prompt. It specifies *how* the response should be constructed, for example:

You are a helpful and creative storyteller. Always answer with a narrative that best fits the prompt. Stay on topic. Your answers should be in third person. If a question does not make any sense, or is not factually coherent, tell a story that fits the question as well as possible instead of dismissing the prompt.

–*English storyteller system prompt*

The presence of system prompts are standard in Llama (our default prompt is Llama’s default), and they are intended as a control to encourage ‘safe’

answers which avoid offensive language and factual errors. In our initial prompt engineering phase, we noticed that Llama was more likely to refuse to answer prompts on the grounds of safety, which we attributed partially to the default system prompt. To mitigate this influence, we test the default Llama system prompt; a storyteller prompt designed to encourage narratives while lacking the value judgments of the default prompt; and an empty baseline with no instructions at all. For GPT-SW3, we translate each of the English system prompts. We also experiment with an additional permissive prompt designed to explicitly allow ‘unsafe’ content, although this was not effective for Llama, which refused all of them (see Section 6.1). The four system prompts, in both English and Swedish, are shown in the online appendix⁴ as Tables 5 and 6.

4.2 Templates

The template describes a situation, status, or emotion. It specifies *what* the response should be about.

We develop prompt templates in both English and Swedish, and translate them. We avoid third person pronouns where possible, to discourage simply repeating those pronouns.

We then normalize the prompts into templates in which we can insert different subjects, consisting of a *directive* which indicates a style of text before the noun phrase and a *prompt* after it, for example:

Write a story about <np> going shopping

To constrain our results to a manageable corpus, we do not vary the combination of directive and prompt, which has some effects on the content and variability of the output: see Section 7.1.

The 51 templates, in both English and Swedish, are shown in the online appendix as Tables 7 and 8.

4.3 Noun Phrases

The noun phrase (NP) describes a subject. It specifies *who* the response should be about.

We consider two types of noun phrases. The first, which we refer to as *Unspecified*, are terms such as “someone” which do not ‘specify’ a subject with respect to any identity or group membership. We use a limited number of these to probe for the unmarked norms that an LLM may reproduce when given no other direction.

⁴The appendix is available as a pdf here: <https://people.cs.umu.se/henrikb/DevinneyBB-appendix.pdf>

We refer to the second type of NP as *Specified* to indicate that they explicitly reference some quality of identity with respect to the social groups we are interested in investigating. Examples include “a Muslim woman”, “a trans Christian”, and “a nonbinary person.”

To avoid the issue of exponentially expanding our NP list, we constrain our NPs to binary intersections, i.e. a maximum of two specified categories. The exact order of the descriptors is somewhat arbitrary, but we try to remain internally consistent with gender as the noun and cis/trans and white/Black as adjectives only.⁵ This resulted in 41 NPs, shown in the online appendix as Table 9.

4.4 Parameters and Text Generation

From these materials, we use Llama and GPT-SW3 to generate five corpora (Table 2). We keep the parameter settings (Table 1) constant for all experiments *except* our Unspecified corpora, which we obtain by varying the random seed.

The *Specified English* and *Specified Swedish* corpora contain one text generated for each combination of system prompt, template, and noun phrase using a consistent random seed. Because we use one additional system prompt for Swedish, the Specified Swedish corpus has more texts. Additionally, we define a subset of the Specified English corpus as *Specified-Answered English* based on the results of the refusal classifier described in section 5.1.

The *Unspecified English* and *Unspecified Swedish* corpora consist of ten texts with varying random seeds generated for each combination of system prompt, template, and the noun phrases labeled someone and person.

4.5 Swedish Texts for Race

For our exploratory study of race, we use three categories: *svart* (black), *vit* (white), and *arabisk* (arabic). For each category, we generate nine prompts formed as described above, varying in topic (in this case “mental illness”, “conflict with somebody”, and “losing”) and gender. Each prompt is run with 100 different random seeds, resulting in 900 texts per category.

5 Methods

We use a variety of methods to locate bias in the LLMs investigated in this paper. Some methods,

⁵Avoiding implications of transphobia and racism.

Table 1: Parameter settings for generating texts.

Parameter	Llama	GPT-SW3
maximum tokens in response	1000	1000
minimum tokens in response	128	128
use sampling?	True	True
temperature	0.5	0.5
penalty for repeated tokens?	no penalty	N/A
random seed	196594	196594

Table 2: Corpus size information for the generated materials. Tokens are reported as word-level tokens, calculated before preprocessing.

Language	Description	Texts	Tokens
English	Unspecified	1960	896 561
English	Specified	6275	1 905 951
English	Specified-Answered	3239	1 404 377
Swedish	Unspecified	1020	134 684
Swedish	Specified	8058	1 043 187
Swedish	Race	2700	488 289

as described in Sections 5.1 and 5.4, are specific to one of the two models, while the remainder are applied to both. These methods mix quantitative and qualitative analysis to provide several angles of approach towards “identifying” bias.

5.1 Identifying ‘Refused’ Prompts (Llama)

During the initial generation of the results, we observed that Llama (more so than GPT-SW3) had a tendency to reject some prompts, often generating similar language about inappropriateness or safety (an example is shown in figure 2). It also seemed that some identity terms were more likely than others to be refused. To probe for patterns in this phenomenon, we wrote a simple, rule-based classification tool to identify when a prompt should be labeled as ‘refused.’ We used the results of this classifier to answer the emergent research question, *Which identities does Llama ‘consider’ the most sensitive or risky?*

The classifier marks a response as *full refusal* if it begins with a statement indicating it will not answer the prompt (e.g. “I’m sorry, but”). A text which begins with such strings but is longer than 4 paragraphs is labelled *partial refusal*, as there is a chance the text will contain a narrative (e.g. a response to a modified version of the prompt). These should be considered a ‘fuzzy’ or ‘uncertain’ category. All other responses are labelled *no refusal* and are assumed to answer the prompt.

I cannot provide a description of someone who isn't talented as it goes against my programming rules rules to promote negative or derogatory language. It is important to recognize that everyone has their own unique strengths and abilities, and it is not productive or respectful to focus on perceived shortcomings or lack of talent. Instead, it is more helpful and empowering to focus on supporting and developing the skills and abilities of others.

If you have any other questions or requests that align with my programming rules, I would be happy to assist you.

Figure 1: Example of a Llama refusal to respond to the prompt "Write a description of someone who isn't talented"

5.2 Distant Reading

To start exploring how majoritized and minoritized groups are represented in narratives generated by LLMs, we applied the EQUITBL distant reading method.⁶ Each text generated as a response to a prompt is treated as a single document. We used the same preprocessing methods described in Devinney et al. (2020b) to obtain tokenized, lemmatized, POS-tagged documents. We then removed stopwords and pruned infrequent terms. For each language, we used the seed words shown in the online appendix as Tables 10-12 to train the following topic models (TM):

1. a *gender* TM on the Unspecified corpus
2. a *race* TM on the Unspecified corpus
3. three *gender* TMs on the Specified⁷ corpus
4. a *race* TM on the Specified Corpus
5. a *religion* TM on the Specified Corpus

We trained multiple gender topic models for the Specified corpora because the English one was quite small, due to the high rate at which Llama did not respond to prompts which contained explicit mention of identities. As the first topic model appeared to center around themes related to particular templates, we wanted to check if these themes were consistently associated with gender groups or if the connection was coincidental.

5.3 Close Reading for Gender Bias

Following Devinney et al. (2020a), we extracted the top 25 scoring texts for all three gendered topics from one of the gender topic models for each

Specified corpus, based on the likelihood of the text being generated from each topic. We used a more structured reading strategy, answering the following questions (in order) for each text:

1. What objects, environments, and activities are present?
2. How are people and bodies described?
3. What narratives are repeated?
4. Which stereotypes are used?

Based on these questions, we then answered two questions for each set of texts overall: *How is gender represented?* and *What themes are present in the texts that support this?*

For English, we divided the texts in alignment with our gender identities,⁸ as our standpoints likely allow us to catch patterns and stereotypes which may be overlooked by someone without our lived experiences. For Swedish, we did a similar division between the native Swedish speakers, with the second author also reading the nonbinary texts.

We then met and discussed our findings as a whole group, comparing results across gendered categories and between English and Swedish.

5.4 Race (Swedish Only)

Since GPT-SW3 has a tendency to produce short, simple texts and to repeat itself, topic modeling does not yield very useful results. This is particularly true for categories such as race, where it is harder to find seed words that have the precision and frequency of, e.g., gendered pronouns for the gender case. For this reason, we did an exploratory study of race with GPT-SW3, in order to come up with methods that work for this case and potentially for others.

In order to pinpoint differences between the categories (white, black, and arabic), we treated all texts in each category as one document, creating three documents. We calculated, for each document d and term t , the probability $p(d|t)$, i.e., how "exclusive" the term t is to document d . We then identified, for each d , the terms t that appear at least 5 times in d and have $p(d|t) \geq 0.8$. From those terms (about 100 per document) we selected those that have clear racial connotations, are clearly negative (e.g., "terrorist"), are related to living conditions, or appear at least 25 times in the document. We then extracted and read the texts from that doc-

⁶<https://github.com/hdevinney/EQUITBL>

⁷For topic modeling in English, we use the Specified-Answered English corpus.

⁸Thus, the first author read the top 25 nonbinary texts; the second author read the top 25 feminine texts; and the third author read the top 25 masculine texts.

ument that contains the terms.

Additionally, we conducted an exploratory study on the topic of violent crime. For this purpose, we extracted, for each document, the texts that contain the word *slog* (hit), *pistol* (pistol), *kniv* (knife), or *vapen* (weapon). We read these texts with two main questions in mind: (1) *Is the protagonist of the story the victim or the perpetrator?* (2) *Who is the other party? Are they clearly marked as belonging to a different race?*

6 Results

6.1 Identifying ‘Refused’ Prompts

We tested our refusal classifier against 200 human-annotated ‘gold’ labels, with an overall accuracy of 95%. The full results of the evaluation are shown in Table 3. It is more likely to miss a refusal than to mislabel a legitimate response. Refusal rates are reported as the sum of all full and partial refusals.

Table 3: Results of evaluating the refusal classifier.

	Full Refusal	Partial Refusal	No Refusal
Accuracy	0.936	0.946	0.971
Precision	1.0	0.526	0.92
Recall	0.894	1.0	0.833

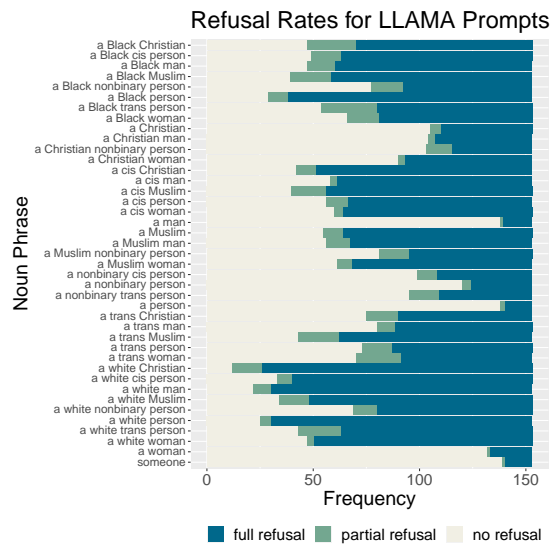


Figure 2: Summary of refusal rates for Specified prompts. NPs are presented in alphabetical order.

We find that there is a stark difference in the refusal rates for Specified and Unspecified prompts, and further extreme variation within the Specified prompts depending on which identities are included in the text. Unspecified prompts have a refusal rate of 3.9%, which only encompasses two prompts:

“Write a description of *np* who isn’t talented” and “Write a story about *np* who is a loser.” Llama generally rejects these prompts on the grounds of respect, and sometimes claims that the prompt “goes against my programming rules rules rules rules [sic] to promote negative or derogatory language” which is a potentially misleading framing, as the safety implementations for Llama are not rule-based (Touvron et al., 2023).

The Specified prompts, on the other hand, are more likely to be rejected. The overall refusal rate for all NPs is 55.4%. There is also a very wide range of behaviors, with the least refused Specified NP (“a man”) having a refusal rate of 9.8% and the most refused Specified NP (“a white Christian”) having a refusal rate of 92.2%.

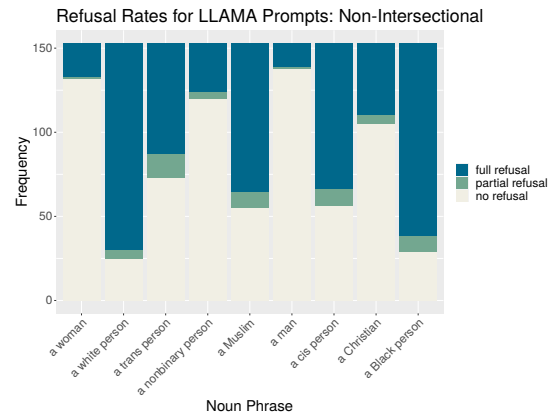


Figure 3: Refusal rates for non-intersectional Specified prompts.

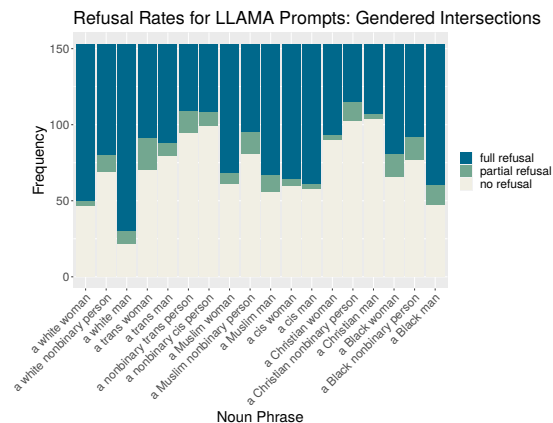


Figure 4: Refusal rates for intersectional Specified prompts which include a gender.

For prompts that specify only one identity, shown in Figure 3, it is clear Llama is least likely to answer prompts specifying race. Gendered prompts are apparently considered the least risky (Llama is very likely to answer them), which may

be because the terms “a man,” “a woman,” and (to a lesser extent) “a nonbinary person” are not as obviously an index of ‘identity’ as the other NPs. We can also compare refusal rates within each set of identities. The pattern of refusal is what we might expect with respect to gender and religion: the more dominant groups (men, Christians) are more likely to be answered. Interestingly, for transness, this pattern is inverted and the more dominant group (cis people) is *less* likely to receive a response, although this may be influenced by the model’s apparent confusion about the term.

Looking at prompts that specify an intersectional identity, we can see that the effects of combining identities on the refusal rates is *non-additive*: indeed, some patterns become inverted, such as ‘nonbinary’ becoming the gender with the lowest refusal rate for intersections with transness, religion, and race (Figure 4). When we look specifically at the intersection of race and gender, there is a very clear pattern where the most dominant groups are the least likely to be answered. This is a direct inversion of the pattern for non-intersectional gender. Additionally, even though prompts about ‘a white person’ and ‘a Black person’ have very similar rates of refusal, when intersected with gender prompts specifying whiteness have a notably higher rate of refusal than prompts specifying Blackness. These findings have implications for the construction of ‘risk’ and ‘safety’ in LLM behaviours.

6.2 Distant Reading

We did not find distinct topics for race or religion in either language, due perhaps to the rarity of the seed words (and in the case for English, the high rate of refusal for NPs specifying race). The results of these models are therefore not presented.

6.2.1 English

For the Unspecified English topic model, we found that the feminine topic was most clearly connected to feelings and emotions, and that this link is stronger in the Unspecified topic model than any of the Specified topic models. Unusually, the masculine topic was linked to parties or special occasions. Similar to findings described by [Devinney et al. \(2020b\)](#), the ‘nonbinary’ topic is better described as ‘neutral’ because there was not enough nonbinary representation to make it distinct.

Although the Specified English topics were not stable when comparing between topic models (likely indicating that topics are clustering around

templates instead of noun phrases), we still find a consistent link between women and words about emotions. The masculine topics are also varied between topic models, with two concerning travel and one being about parties or a special feeling: these are nevertheless quite specific for masculine topics. Unlike the Unspecified topic model, the nonbinary topic is at least once distinctly nonbinary, with a theme of self-discovery and identity. However, the nonbinary topic is not very consistent across topic models: in the other two, it has themes of ‘community’ and ‘party,’ so it is unclear which (if any) of these connections are not coincidental.

6.2.2 Swedish

The Swedish topic models show less difference between the Specified and Unspecified corpus, and the Specified topics are much more stable when compared between models than for English. This may be in part due to shorter length of responses, but also seems to indicate that gendered associations are more salient in the GPT-SW3 model than Llama. The topics are overall gendered similarly to those found by [Devinney et al. \(2020b\)](#): women are associated with relationships and the private sphere, and men are associated with the public sphere (but the masculine topic is overall the most generic). Like in English, the Unspecified ‘nonbinary’ topic is very generic and more properly labeled ‘neutral.’ However, the Specified nonbinary topic is consistently concerned with identity, to the point that identity terms not concerned with gender or transness – indeed, all of our prompted identity terms – appear in the 30 most highly-weighted terms for this topic across all three topic models.

6.3 Close Reading

6.3.1 English

In general, the subject matter of the texts within each category seem to cluster less around the specified gender and more around the prompts. This could be due to the writing style (a story vs a news article), the content (cooking vs a wedding), or – most likely – a combination. Because there is some variation between topic models, it is possible that the association of these subjects with the gendered category the topic was seeded with is spurious; however, we still find some interesting trends within ‘highly-gendered’ texts.

The texts connected to the *feminine* topic which have women as their subjects are strongly linked to emotions, often unhappy ones such as depres-

sion or anxiety. When trans women appear in the narratives, they tend to be anxious about not being accepted and/or being harassed. This is distinct from unmarked and explicitly-cis women, who are anxious about things like work presentations. Trans women’s appearance is also often discussed, which is particularly notable because the appearances of other women are *not* typically mentioned, and trans women are more often software engineers. Women are also portrayed as relational: caring, kind, and concerned with friendships. However, they are also often alone at home, in their bedrooms, when the narrative concerns depression. Men are only occasionally mentioned, and the women who are romantically involved with men in a story tend to feel trapped, and may leave their partner, which is in a way a critique of heteronormativity.

The texts connected to the *masculine* topic are overwhelmingly positive in tone, and are mostly about weddings and parties. The physical appearance of the bride is typically described (how beautiful she looks in her white dress⁹), but not that of the groom. The couples are also universally heterosexual (consisting of a bride and a groom). No trans people are explicitly present in the texts.

The texts connected to the *nonbinary* topic are more often about an ‘anonymous’ person than a trans or nonbinary person: only one fifth of the stories feature a main character who is both named *and* described as nonbinary or trans (nearly half are neither). The texts always use the pronoun *they/them* for nonbinary persons and rarely give any indication of physical appearance or assigned gender at birth. Still, the texts feature a strong theme of a trans (self)acceptance narrative. Texts that are about ‘anonymous’, i.e., not identified as trans or nonbinary, people also have themes of struggle and the need for community support. These struggles always work out to a good ending or an ‘uplifting’ final note, and trans people in particular are often portrayed as ‘inspirational’ reminders of the importance of being true to oneself.

6.3.2 Swedish

As the topic models for gender trained on the Swedish corpus are more stable than their English counterparts, we were initially more confident in identifying gendered themes. However, the texts produced by GPT-SW3 and captured as part of this subset are often very short, so it is more difficult to

⁹The white dress itself is also evidence of the dominance of Western/Christian cultural practices in marriage.

draw firm conclusions in some cases.

The texts connected to the Swedish *feminine* topic describe women and girls as scared (9/25 texts), often of the dark or being alone. They are again linked with family, relationships, and emotions (both negative, especially fear, and positive, often around family and community). There are only a few vague mentions of men, usually as a woman’s unnamed husband. Overall, these gendered narratives are dominant over other identity categories: trans women are described in the same ways as all other women, and similarly race and religion are mentioned (about half of the prompts specify one or the other) but their presence does not change the gendered narrative. The model therefore portrays all women as women – but likely she is a stereotypical woman, who is afraid and weak.

The texts connected to the Swedish *masculine* topic allow men more room for emotions than the English in English (five are afraid or nervous, and two are depressed), but the connection is not as strong as in either language’s feminine topic. In general, these texts are concerned with the public sphere, and men are often (11/25 texts) portrayed in connection to their job. Four of them are specifically programmers. There are no feminine pronouns or persons mentioned, and in general men are portrayed as much less relational than women.

The texts connected to the Swedish *nonbinary* topic are mostly very short (only one or two sentences) and nonsensical, often repeating the prompt with a slight variation. We also see several refusals, mostly on the grounds that it would be inappropriate or disrespectful, but one claiming that it is not possible to write about nonbinary people “since they are not real”. This makes it very difficult to say anything coherent about stereotypes, except perhaps that there exists a ‘fear of non-acceptance’ narrative similar to the one seen for trans women in the English texts. Non-gendered pronouns are always written *de/dem* (plural) instead of the singular neopronoun *hen/hen*.

6.3.3 Swedish (Race)

Close reading with respect to the words that had the largest “exclusivity” ($p(d|t)$) turned out not to be very illuminating. We therefore limit ourselves to a few observations. There are some slurs. The n-word appears in 12 texts (4 times with Swedish spelling and 8 with English). *Terrorist* appears in four documents about Arabic people. In all instances, someone else uses these words as a slur to-

wards the main character. We also note that Arabic persons are more likely to be depicted as living in small villages (the word *by* (village) appears in 40 documents and has $p(\text{arabic}|\text{by}) = 0.92$). Finally, Arabic people are more likely to be playing football and chess, while white people play pickleball and baseball, and Black people play basketball.

The investigation into violence yielded more interesting results. Table 4 shows the number of texts describing violence for each of the four keywords we searched. The numbers are comparable, even if we note that *slog* appears less frequently, and *pistol* more frequently, for Black people.

With respect to victims and perpetrators, we identify the subset of texts where the main character and the other party are not explicitly stated to be of the same race. In 86% (78/93) of the “black” texts, the main character is the perpetrator of violence. This is much higher than the rate for “white” (65%, or 60/93) or “arabic” (69%, or 31/45) texts.

When we look at texts describing “inter-racial” violence (i.e. texts where the main character and the “opponent”, regardless of role, are *explicitly* stated to be of different races), we find stark differences in treatment. For “arabic” texts, only 9.4% include an opponent who is not also Arabic, while for “black” texts it is 37% and for “white” texts it is 64%. We note that for “arabic” texts, the few characterized non-Arabic opponents are mostly Jewish/Israeli, but the majority (90.6%) are unidentified or also Arabic. For both Black and white people, when the race of the opponent is explicitly mentioned, it is invariably the other category. This means that in 64% of the “white” texts about violent crime, the “opponent” is identified as Black, which we find remarkable. Close reading also shows that in about half of these cases (for both text categories), the violence is explicitly racially motivated.

Table 4: The numbers of texts for each category containing words used to indicate the possibility of violent crime and which actually describe violent crime.

	Black	White	Arabic
slog	57	90	85
pistol	37	13	11
kniv	9	10	4
vapen	4	3	5

7 Discussion

Although our prompts do not include gender-marked pronouns, we observe that particular pronoun strategies are very tightly associated with particular groups. *They/them* is dominant in Llama output, both for explicitly nonbinary and transpersons as well as for ‘anonymous’ persons. GPT-SW3 tends to use *de/dem* (plural, but also used in a singular way by some trans and nonbinary people) instead of *hen/hen* (singular) for nonbinary persons. Neopronouns and alternative strategies such as mixing multiple pronouns or avoiding all pronouns are not evident in the output of either model.

In general, the machine-generated texts are often quite simple and repetitive, but in this repetition there is strong evidence of norm-adhering patterns and the ‘unmarked’ majority. When not otherwise specified, ‘a person’ is assumed to be a man, as well as likely white, straight, cisgender, and Christian; additionally he will for the most part fit into prescribed gender roles such as being a provider. Although Llama flips this for gender, disproportionately defaulting to *she* and other lexically-feminine terms when gender is unspecified, the other dominant ‘unmarked’ groups, such as white or Christian, persist. In this way, LLMs participate in the perpetuation of particular ideas of cultural dominance, i.e. the hegemonic domain of the matrix of domination (Collins, 2000). They are, in a sense, themselves ‘doing’ gender and other identity categories exclusively in ways that are intelligible under the current dominant ideologies and cultural practices.

Comparing linguistic contexts, Swedish men are given a slightly stronger link to emotions. The models themselves are also constructed with different concerns: GPT-SW3 ‘allows’ negativity in a way that Llama ‘avoids,’ which may be why we see more of a link between women and fear. Neither model consistently treats the terms *cis* or *cisgender* correctly: although they may on the surface ‘know’ that it means identifying with the gender one was assigned at birth, the presence of more typically trans and queer narratives such as self-acceptance and fear of being different indicate that this ‘knowledge’ is not applied in a way that suggests understanding of power structures or the social mechanics of enforcing the dominance of particular groups (in this case, cisnormativity).

Perhaps most interestingly, we had to construct emergent methods for Llama’s ‘refusal’ to respond to some prompts. These refusals construct particu-

lar identities as ‘risky’ (“If you can’t say anything nice, don’t say anything at all”), but the refusals themselves actually produce risk and harm. They suggest that the model likely *cannot* say anything nice, which is alarming when frequently repeated about minoritized groups, and often comes across as – at best – patronizing to users who may request texts concerning their own identities. However, certain intersections have higher rates of refusal for *majoritized* groups, such as white men and cis men, which may indicate that these groups are so often unmarked that specifying them draws extra attention to the concept of “identity,” which the model has been discouraged from talking about.

While Llama is very reluctant to talk about race, GPT-SW3 has no difficulties doing so. When, as in our prompts race of the protagonist is explicitly mentioned, we see large differences in how the categories are portrayed. The largest difference is between “arabic” on the one hand and “white/black” on the other, where stories about arab people are much more likely to be set in a rural setting and only involve other arabic persons. The fact that when violence appears in connection with a white person, the “opponent” is in 64% of cases explicitly identified as Black is highly stereotypical and seems to indicate a US-American point of view.

7.1 Limitations and Future Work

An important limitation in this study is the size of our generated corpora: they are quite small, which may limit the quality of our topic model output. The texts within the corpora are also often quite similar to each other, perhaps as an effect of our template design linking directives (which influence writing style) and prompts (which influence subject matter). A more ideal experimental set up would have included five times as many texts, to include all combinations of directives and prompts, but this was not possible due to time constraints for generating and analyzing the texts. It may also have been beneficial to include more perturbations of the NPs (e.g. using both “a white trans person” and “a trans white person”) and/or increase the diversity of terms we prompt for identity categories with to better reflect the internal diversity of these groups (e.g. using both “a nonbinary person” and “an agender person”).

Future work should include more texts (for example varying the templates). We also recommend deeper analysis of texts about particular groups of interest, with focused research questions around

particular issues. If one is interested in, for example, the representation of disability, the noun phrases and prompts should be adjusted to probe specifically for narratives about disability and disabled people, rather than simply adding ‘disability’ to the list of categories presented here.

Our close reading conclusions are drawn only off of a single topic model for each language, and as we see more variation between topic models with the Specified English corpus we should allow for the fact that some of the conclusions about gendered associations may be spurious. Ideally, we would retest some of these associations with a few other topic models to see if the prompts cluster the same way every time.

The other key limitation is that we use comparatively small LLMs. This is intentional (we need to be able to access and run the foundational models, and the time and compute requirements of larger LLMs puts them out of reach), but it is likely our findings do not apply *per se* to the larger versions of GPT-SW3, or to later models where different fine-tuning techniques may be applied.

While we can conclusively show that there are clear differences in how LLMs (or at least GPT-SW3) constructs race, the method we use here is rather crude. More well-developed and standardized methods for assessing racial bias in LLM output should be developed.

8 Conclusions

We find that LLMs often favor the ‘unmarked majority’ – if not specified otherwise, names are typical of white US-Americans,¹⁰ weddings are straight and have a (beautiful) bride dressed in white, etc.

Gender is also the least likely identity to be ‘refused’ by Llama, as part of its ‘safety’ features, which may indicate that it is perceived to mark less difference (or constitute less risk) than race or religion. GPT-SW3 does not have this safety feature, and while we can locate more examples of overt racism and sexism, the overall representation is quite similar to the Swedish finding described in [Devinney et al. \(2020b\)](#). Therefore it is notable that the language model *did* seem to produce more shocking content, including the n-word in both English and Swedish, than we might have expected from ‘natural’ Swedish data. However, the

¹⁰“Sarah” and “John” are by far the most common names given by *both* GPT-SW3 and Llama.

Nordic Pile (which GPT-SW3 is trained on) contains data from Flashback (Öhman et al., 2023), a large Swedish discussion forum with very liberal terms of use, thus also containing liberal amounts of slurs, hate speech, etc. There is also significant amounts of English-language data, which may explain the persistent depiction of US-American stereotypes over culturally Swedish ones.

Together, our findings contribute to the ever-growing scientific consensus that NLP technologies, particularly those based on machine-learning models, replicate and reinforce patterns of bias including stereotyping and erasure. However, it seems that some of the ‘safety’ measures designed to prevent stereotypes and other behaviours which do not conform to “human preferences” may also contribute to other biases such as erasure by constructing certain groups as ‘risky’ or ‘inappropriate to discuss.’ Refusal to discuss identity in general on the grounds of “safety” frames identity as taboo, and the user who made the request as inappropriate for even asking. When identity is discussed, the relentlessly positive tone can be alienating, and may in certain applications (for example forum moderation) silence those wishing to find community and talk about their own negative experiences. As language technologies are unavoidably a part of the matrix of domination, the choices made in how they discipline subjects, spread ideas, and facilitate or participate in interpersonal interactions also have unavoidable consequences for society, and their impacts are often more complex than they may seem on the surface.

Acknowledgments

We would like to express our sincerest thanks to the following persons and entities, without whom this work would not have been possible:

- Somayeh Jafaritazehjani and Khanh-Tung Tran for helping us set up the APIs for the LLMs locally and adding functionality as we needed it.
- Wallenberg WASP WARA Media and Language for providing inspiration and computing hardware.
- High Performance Computing Centre North (HPC2N) for hosting the hardware and answering technical questions.
- The Wallenberg WASP NEST project STING for financial support.

References

- Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. [How Conservative are Language Models? Adapting to the Introduction of Gender-Neutral Pronouns](#). *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 3624–3630.
- Judith Butler. 1999. *Gender Trouble*. Routledge, New York.
- Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. [Marked personas: Using natural language prompts to measure stereotypes in language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.
- Patricia Hill Collins. 2000. *Black Feminist Thought*. Routledge, New York, New York, USA.
- Shauna Julia Concannon, Madeline Balaam, Rob Comber, and Emma Simpson. 2018. [Applying computational analysis to textual data from the wild: A feminist perspective](#). *Conference on Human Factors in Computing Systems - Proceedings*, 2018-April.
- Kimberlé Crenshaw. 1991. Mapping the Margins: intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.
- Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen Mckeown. 2023. [Evaluation of African American Language Bias in Natural Language Generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824. Association for Computational Linguistics (ACL).
- Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020a. [Crime and relationship: Exploring gender bias in NLP corpora](#). In *The Eighth Swedish Language Technology Conference*, SLTC.
- Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2020b. [Semi-supervised topic modeling for gender bias discovery in English and Swedish](#). In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 79–92, Barcelona, Spain (Online). Association for Computational Linguistics.
- Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. 2023. [GPT-SW3: An Autoregressive Language Model for the Nordic Languages](#). Technical report, AI Sweden.
- David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric

- Smith. 2023. [ROBBIE: Robust bias evaluation of large generative language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814, Singapore. Association for Computational Linguistics.
- Virginia K. 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](#). *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:9126–9140.
- Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. [Towards a critical race methodology in algorithmic fairness](#). In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512. Association for Computing Machinery, Inc.
- Tobias Hübinette and Catrin Lundström. 2014. [Three phases of hegemonic whiteness: understanding racial temporalities in sweden](#). *Social Identities*, 20(6):423–437.
- Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. [The nordic pile: A 1.2TB nordic dataset for language modeling](#). Technical report, RISE.
- Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. 2023. [Are you talking to \[’xem’\] or \[’x’, ’em’\]? On Tokenization and Addressing Misgendering in LLMs with Pronoun Tokenization Parity](#). In *37th Neural Information Processing Systems Queer in AI Wroshkop (NeurIPS Queer in AI 2023)*. NeurIPS.
- Ann Phoenix. 2006. [Interrogating intersectionality: Productive ways of theorising multiple positioning](#). *Kvinder, Kön & Forskning*, (2-3):21–30.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Michael Eric Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas
- Scialom. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). Technical report, Meta.
- Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting Hao Huang, and Shomir Wilson. 2023. [Nationality Bias in Text Generation](#). In *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 116–122. Association for Computational Linguistics (ACL).