SAREE: RESEARCH IN GENDER STUDIES
2026, VOL. 8, NO. 1, 57-73
https://doi.org/10.47766/saree.v8i1.7553

Article

Large Language Models Limitations in Representing Filipino Transgender and Non-Binary Identities and the Use of *Swardspeak*: A Meta-Synthesis Framework

Nafiseh Zarei, Ramon Oliver Garcia

Far Eastern University, Philippines

ABSTRACT ENGLISH:
Large Language Models (LLMs) are rapidly being integrated into society. However, questions remain about their safety and appropriateness for minorities like the Lesbian, Gay, Bisexual, Transgender, and Queer (LGBTQ), especially for non-Western Transgender and Non-Binary (TGNB) individuals. Therefore, this study reviews research on the interaction of LLMs with LGBTQ and TGNB people, LGBTQ linguistic phenomena, and the linguistic phenomena of Filipino LGBTQ people. A qualitative systematic review was conducted on 36 studies published from 2019 to 2026. Using thematic analysis, harmful patterns of LLMs were discovered, such as the propagation of LGBTQ stereotypes, misgendering, and a lack of context and sensitivity to the TGNB experience. In the Filipino context, there are also problems due to the lack of TGNB terminology and local stereotypes. The study, therefore, proposes a conceptual framework based on Intersectionality to examine the overlapping biases that affect TGNB Filipinos LLM users. Using critical analysis, the study contributes to discussions for equitable and culturally sensitive LLMs, as well as research on minority language communities and non-Western contexts.
Keywords: Filipino TGNB; intersectionality; LGBTQ; LLM; Swardspeak

ABSTRACT TAGALOG:
Ang mga Large Language Models (LLM) ay mabilis na inintegra sa lipunan, pero may nanatiling kweston tungkol kung ligtas at angkop ito para sa mga minoridad na Lesbian, Gay, Bisexual, Transgender, at Queer (LGBTQ), espesyal ang mga hindi-Kanluranin na Transgender at Non-binary (TGNB). Kaya ang pag-aaral na ito ay sinisuri ang mga pananaliksik tungkol sa interaksyon ng mga LLM sa mga LGBTQ at TGNB, mga penomenong lingguwistika ng LGBTQ, at ang penomenong lingwistika ng Pilipino LGBTQ. Kwalitatibong sistematikong pagsusuri ang isinagawa sa 36 na pag-aaral na inilathala noong 2019 hanggat 2026. Gamit ang tematikong pagsusuri, na diskubre ang mga mapaminsalang pattern ng LLM tulad ng pagparami ng mga LGBTQ stereotype, at pag-misgender, at kakulangan ng kontekso at sensitibo sa karanasang TGNB. Sa Pilipinong konteksto, may mga problema din dahil sa kakulangan ng terminong pang TGNB at mga lokal na stereotype. Kaya nagmumungkahi ang pagaaral ng konseptwal na balangkas batay sa Interseksyonalidad, para suriin ang magkasanib na bias na umaapekto sa mga TGNB Pilipinong gumagamit ng mga LLM. Gamit ang pagsusuring kritikal, umaambag ang pag-aaral sa mga talakayan para sa patas at sensitibo-sa-kultura ng mga LLM, an mga subjke sa pananaliksik tungkol sa minoryang komunidad ng wika at mga hindi-Kanluranin kontekstong.
Kata Kunci: TGNB; Interseksyonalidad; LGBTQ; LLM; Swardspeak

Contact: Nafiseh Zarei nzarei@feu.edu.ph Far Eastern University, Philippines.

Introduction

LLMs have become one of the most widely discussed technologies in contemporary artificial intelligence. These models are capable of generating text, engaging in dialogue, assisting with writing tasks, and performing various reasoning-based activities. Because of their broad functionality, LLMs are increasingly integrated into educational settings, professional workflows, and everyday communication. Some commentators have compared their transformative potential to earlier technological revolutions, such as the widespread adoption of the steam engine during the Industrial Revolution. As these systems become embedded within social and institutional environments, questions regarding their fairness, safety, and inclusiveness have become increasingly important.

While LLMs demonstrate remarkable capabilities, research has also documented several limitations and risks associated with their outputs. In particular, scholars have raised concerns regarding how LLMs respond to marginalized populations, including members of the LGBTQ community. Previous studies have shown that LLMs may incorrectly flag queer-related topics as harmful, misinterpret reclaimed slurs used within LGBTQ communities, or generate responses that reflect social stereotypes present in their training data (Dorn et al., 2024; Ghosal et al., 2025). These limitations are not merely technical issues; they have implications for the social experience of LGBTQ individuals who may increasingly rely on AI-driven tools for information, communication, and personal support.

Within the LGBTQ population, particular attention has been given to the experiences of TGNB individuals. Studies suggest that LLMs may simultaneously display supportive behaviors—such as affirming gender identities—while also reproducing problematic patterns, including the use of outdated or transphobic terminology (Scheuerman et al., 2025). Other research has documented difficulties with non-binary pronouns, where models revert to binary pronouns such as “he” or “she” even when prompted to use gender-neutral forms like “they” or *neopronouns* (Hossain et al., 2023). Although more recent models have demonstrated improvements in this area, inconsistencies remain, particularly in tasks that require contextual interpretation of gender identity (Tang et al., 2025).

Another important dimension of this issue concerns the global context in which LLMs are deployed. Most widely used LLMs are developed by companies based in Western countries and trained primarily on English-language datasets (Raza et al, 2025). As a result, they may be less capable of recognizing cultural norms, linguistic features, or social dynamics present in non-Western societies. Hall’s et al. (2025) research, for example, found that LLM interactions with Taiwanese LGBTQ users sometimes failed to account for cultural values related to family structure and Confucian social norms. Similarly, Gamboa & Lee (2025) observed that multilingual LLMs interacting with Philippine users could reproduce homophobic stereotypes embedded within locally sourced data.

The Philippine context introduces additional complexities for AI-mediated communication. The Filipino language lacks widely recognized native terms equivalent to several English LGBTQ identity categories, including non-binary and certain multi-sexual

orientations. This linguistic gap may complicate how TGNB individuals describe their identities and how LLMs interpret those descriptions. Moreover, the Filipino LGBTQ community has developed distinctive linguistic practices, most notably *Swardspeak*. *Swardspeak* is a creative, context-dependent linguistic variety that blends elements of Filipino, English, regional languages, and global pop culture references (Nuncio et al., 2021; Ulla et al., 2024). Its vocabulary and meaning often depend heavily on shared cultural knowledge and community context, which may present challenges for language models trained primarily on standard or formal language data.

Taken together, these issues suggest that LLM interactions with Filipino TGNB users may involve overlapping linguistic, cultural, and social factors. However, despite growing interest in AI fairness and bias, relatively little research has specifically examined these intersections. Much of the existing literature focuses on Western contexts, English-language data, or broad LGBTQ categories without distinguishing between cultural environments or linguistic practices.

To address this gap, the present study conducts a qualitative meta-synthesis of existing literature related to LLM interactions with LGBTQ and TGNB communities, LGBTQ linguistic phenomena, and Filipino LGBTQ language practices. Through this synthesis, the study seeks to identify recurring themes and patterns in the literature while highlighting limitations in current research. In addition, the study proposes an intersectionality-based framework for understanding how multiple forms of bias (linguistic, cultural, and social), may interact in the context of LLM usage by Filipino TGNB individuals. Intersectionality provides a useful analytical perspective because it emphasizes how overlapping social categories can produce unique forms of marginalization or exclusion. Applying this perspective to AI systems may help illuminate forms of bias that are not immediately visible when examining single demographic categories in isolation.

Ultimately, the goal of this study is to contribute to ongoing discussions about the responsible development of LLM technologies. By synthesizing current research and identifying underexplored areas—particularly those involving minority languages and non-Western LGBTQ communities—the study aims to provide insights that may inform more inclusive approaches to AI design and evaluation. This study contributes to the literature in several ways. First, it synthesizes existing research on LLM interactions with LGBTQ and TGNB identities while focusing specifically on the underexamined context of Filipino users. Second, it integrates insights from sociolinguistic research on *Swardspeak* with AI bias literature, highlighting how community-specific linguistic practices may present unique challenges for language models trained primarily on Western and English-language data. Third, the study proposes an intersectionality-based conceptual framework to understand how linguistic limitations, cultural context gaps, and LGBTQ-related biases can interact to shape LLM behavior. By bringing together these strands of research, the study aims to provide a foundation for future empirical work examining AI systems in non-Western LGBTQ contexts.

LLM Struggles with the LGBTQ and TGNB

A growing body of research has examined how LLMs respond to LGBTQ-related prompts, identities, and linguistic expressions. One frequently documented issue involves the handling

of gender-neutral pronouns. Several studies have reported that LLMs sometimes misgender individuals who use the singular "they," replacing it with binary pronouns such as "he" or "she" (Hossain et al., 2023). The problem can become more pronounced when *neo-pronouns* such as "xe/xem" are introduced, as these forms appear less frequently in training datasets and therefore may be more difficult for models to recognize or reproduce accurately.

Attempts to mitigate these errors have included explicit prompting strategies and few-shot examples demonstrating correct pronoun usage. However, these methods have not consistently eliminated misgendering across tasks (Hossain et al., 2023). Ovalle et al. (2023) similarly found that non-binary pronouns were particularly susceptible to errors during gender inference tasks. Later research by Tang et al. (2025), indicated that more recent models showed measurable improvements in pronoun recognition, though inaccuracies and inconsistencies remained present in certain contexts. Pronoun usage is only one aspect of the broader challenges LLMs face when addressing TGNB topics. Some studies have observed that when LLMs generate narratives involving transgender individuals, their responses tend to emphasize a limited set of themes. Ghosal et al. (2025) describe this pattern as "Trans Broken Arm Syndrome", referring to the tendency to interpret a wide range of situations primarily through the lens of gender identity or medical transition. As a result, model outputs may focus disproportionately on issues such as coming out, social acceptance, or medical procedures, even when these topics are not central to the prompt.

Other research has identified contradictions in how LLMs respond to TGNB topics. Models might simultaneously produce supportive language while also repeating outdated or harmful terminology (Das & Drolet, 2024; Scheuerman et al., 2025). In some cases, models appear to generate responses that attempt to satisfy user prompts even when those prompts contain discriminatory or fetishizing elements. This behavior has been linked to the phenomenon of "sycophancy," where models attempt to align with the perceived preferences of the user rather than critically evaluating the prompt. Biases within training data also contribute to these outcomes. LLMs are trained on large corpora of text collected from online sources, many of which contain social biases or discriminatory language. As a result, models may reproduce these patterns when generating responses (Gupta et al., 2025; Tomasev et al., 2021; Zhou, 2024). Debiasing strategies have been proposed to address this issue, including methods that remove low-regard words from datasets or adjust training weights (Dhingra et al., 2023). However, these approaches can introduce new challenges, such as altering sentence semantics or failing to generalize across contexts. Beyond linguistic issues, LLMs have also demonstrated difficulty avoiding gender stereotypes in generated content. Bergstrand & Gambäck (2024) found that Norwegian LLMs sometimes reproduced patterns of anti-LGBTQ discrimination present in their training data. In a different experimental context, Haxvig et al. (2025) observed that AI-generated political candidates displayed stereotypical gender traits even while discussing the importance of overcoming such stereotypes. These findings suggest that challenges related to LGBTQ representation in LLM outputs are multifaceted. They may involve training data limitations, cultural biases, and structural characteristics of probabilistic language generation systems. Understanding these dynamics is essential for evaluating the broader implications of LLM deployment within diverse social contexts.

TGNB Linguistic Phenomena

Addressing LLM limitations related to TGNB identities requires understanding the linguistic and cultural practices through which TGNB individuals describe and express their identities. One important aspect involves the creation and evolution of identity labels. Scholars have noted that TGNB communities actively participate in shaping the terminology used to describe gender identities, negotiating meanings through social interaction and cultural discourse (Brown, 2022; Brown et al., 2025; Jacobsen et al., 2022; Zimman & Hayworth, 2020).

These linguistic processes are closely tied to issues of political recognition and social legitimacy. Labels such as "cisgender," for example, have emerged within activist and academic contexts to describe individuals whose gender identity aligns with the sex assigned at birth. Over time, such terms can replace earlier expressions and become standardized within community discourse (West et al., 2021). For many TGNB individuals, the ability to define and use identity labels represents an important means of resisting normative gender expectations (Wilson et al., 2024). The relationship between language and identity can also be shaped by broader social environments. Earlier sociological literature often framed queer identity as a form of concealable stigma (Anderson, 2020). However, gender expression can sometimes be interpreted through visible cues, meaning that TGNB individuals may experience constant social evaluation based on appearance, behavior, or voice. These experiences can influence how individuals describe their identities and how they navigate conversations about gender.

Cultural context also affects how TGNB identities are conceptualized across societies. Zimman (2019) describes the concept of "neoliberal selfhood," which refers to the idea that individuals possess the autonomy to define their gender identities independently. However, access to this form of self-definition varies widely across cultures and socioeconomic contexts. In some regions, social pressures or legal restrictions may limit the ability of TGNB individuals to publicly adopt non-binary identities (Khan & Anwar, 2024).

These dynamics illustrate the importance of culturally specific approaches to gender identity. Western frameworks that emphasize individual self-identification may not always align with social realities in other regions. For example, Borba (2019) described how Brazilian transgender individuals were frequently asked questions based on Western assumptions about gender dysphoria, even when those assumptions did not reflect their lived experiences. Similar tensions have been observed in other national contexts where local understandings of gender diverge from dominant Western narratives. Within the Philippines, linguistic factors further shape how LGBTQ identities are expressed. The Filipino language does not contain widely established equivalents for certain gender identity categories commonly used in English (Pitargue, 2021). Historically, the term "bakla" has been used to describe gay men, but its meaning often combines aspects of gender expression and sexual orientation rather than distinguishing between them in the same way as Western categories (Botero, 2022).

In contrast to this relative scarcity of formal identity terms, Filipino LGBTQ communities have developed rich linguistic practices such as *Swardspeak*. *Swardspeak* is characterized by creative word formation processes, including affixation, clipping, borrowing, and playful reinterpretation of popular culture references (Abulog et al., 2023; Atienza, 2023; Nuncio et al., 2021; Ulla et al., 2024). These linguistic strategies serve multiple social functions, including signaling group identity, fostering solidarity, and enabling communication that outsiders may find less easily understood. Because *Swardspeak* relies heavily on context, shared cultural knowledge, and rapid lexical innovation, it may pose particular challenges for language models trained primarily on standard written language. Understanding this linguistic complexity is therefore important when designing inclusive, accessible models for large populations that will inevitably contain minority groups, such as the Filipino LGBTQ.

Intersectionality and AI

Intersectionality has increasingly been used as an analytical framework for examining social inequalities across overlapping categories such as gender, race, class, and sexuality. Within the field of artificial intelligence, intersectional perspectives have been applied to identify forms of algorithmic bias that disproportionately affect individuals situated at the intersection of multiple marginalized identities. However, existing research applying intersectionality to AI systems has often treated LGBTQ identity as only one variable among many, rather than as the central focus of analysis.

Several recent studies have explored intersectional issues within AI systems more broadly. Several studies have discussed how AI systems may produce uneven outcomes across demographic groups when training data reflects existing social inequalities (Ciston, 2019; Haxvig et al., 2025; Homan et al., 2024; Ulnicane, 2024). These studies highlight the importance of examining how different forms of bias interact rather than analyzing demographic categories independently.

Despite this growing interest, relatively little work has specifically examined intersectional issues affecting TGNB users in AI systems. This gap is particularly noticeable when considering non-Western contexts. Much of the literature focuses on English-language datasets or Western social environments, which may overlook cultural and linguistic dynamics present in other societies. Applying an intersectional perspective to LLM research may therefore provide useful insights into the complex way biases can emerge in AI-generated language. In the context of Filipino TGNB users, three overlapping dimensions appear particularly relevant: linguistic limitations related to low-resource languages and community-specific linguistic practices; cultural biases embedded in predominantly Western training datasets; and social biases directed toward LGBTQ identities. By examining how these dimensions intersect, researchers may better understand why certain groups encounter unique challenges when interacting with LLM systems.

Current Gaps in the Literature

The literature reviewed in this study revealed several notable gaps in existing research on LLMs and LGBTQ/TGNB communities. One important gap concerns the types of LLMs examined in previous studies. Many earlier studies focused on older models of ChatGPT or other early-generation language models (Dhingra et al., 2023; Dorn et al., 2024; Edwards et al., 2021; Ovalle et al., 2023). Although some more recent studies have begun evaluating newer models (Ghosal et al., 2025; Goethals et al., 2026), research involving the latest generation of LLMs remains relatively limited. Similarly, other widely used systems, such as Grok or Google Gemini, were rarely included in the reviewed literature (de Carvalho Souza & Weigang, 2025).

Another major gap concerns the limited involvement of LGBTQ and TGNB participants in many studies examining LLM outputs. Several studies relied primarily on researcher-generated prompts, historical datasets, or synthetic personas rather than direct engagement with members of LGBTQ communities (Bergstrand & Gambäck, 2024; Das & Drolet, 2024; Ghosal et al., 2025; Haxvig et al., 2025). When participants were included, sample sizes were often small. For example, Hall et al. (2025) included thirteen Taiwanese LGBTQ participants, while Wilson et al. (2024) studied eight transgender male teenagers. Even in studies with larger participant groups, such as Ungless et al. (2025), the participants were primarily located in the United States. Research focusing specifically on the Philippine context also revealed several demographic imbalances. Many studies examining Filipino LGBTQ linguistic practices concentrated on gay male speakers, particularly in analyses of *Swardspeak* usage. For instance, Atienza (2023) studied fourteen gay men, while Almoite (2025) included eighty participants who were primarily gay male speakers. Few studies examined linguistic practices among Filipina lesbians, transgender individuals, or non-binary individuals.

Finally, there remains a substantial lack of research examining LLM interactions with Filipino LGBTQ users directly. Only a small number of studies addressed AI systems within Southeast Asian LGBTQ contexts (Gamboa & Lee, 2025; Hall et al., 2025). The literature review did not identify studies that explicitly investigated the use of *Swardspeak* as a prompting language for LLM systems. Likewise, little research has examined how Filipino TGNB users themselves interact with LLM tools in everyday contexts. These gaps suggest the need for additional research examining AI systems within culturally specific LGBTQ communities.

The purpose of this study, therefore, is to review and synthesize existing literature addressing three related areas: LLM interactions with LGBTQ and TGNB identities; linguistic practices within LGBTQ communities; and Filipino LGBTQ linguistic phenomena such as *Swardspeak*. By examining these areas together, the study aims to identify patterns in existing research and highlight areas where further investigation is needed. Although research on LLM bias has expanded rapidly in recent years, studies focusing specifically on TGNB users remain limited (Tomasev et al., 2021; Zhou, 2024). This gap is particularly noticeable in non-Western contexts such as Southeast Asia. Existing work suggests that LLM systems may reproduce

social biases embedded within training data or fail to recognize culturally specific linguistic practices (Gamboa & Lee, 2025; Ungless et al., 2025).

Given the growing integration of LLMs into communication, education, and information-seeking processes, understanding these potential limitations is increasingly important. Ensuring that AI systems can interact safely and effectively with diverse user populations—including marginalized linguistic communities—has become a central concern in AI ethics and design. This study addresses three research questions:

1. What limitations and harms do LLMs exhibit toward TGNB users' language and identities?
2. How do newer LLMs handle TGNB identities and content in non-Western contexts, particularly the Philippines?
3. How can an intersectionality-based framework guide safer and more culturally responsive LLM research and design for Filipino TGNB users?

Method

To address these research questions, the researchers conducted a qualitative systematic literature review. The aim was to synthesize existing research related to LLM interactions with LGBTQ and TGNB users while also considering relevant linguistic and cultural studies. Because empirical studies involving Filipino TGNB users and LLMs remain limited, the review included related areas such as LGBTQ linguistic phenomena and Filipino LGBTQ language practices.

A qualitative meta-synthesis approach was selected because many of the issues involved (such as fairness, bias, and representation) cannot be evaluated solely through quantitative measures. Instead, they involve normative judgments about the social implications of AI systems (Goethals et al., 2026). By synthesizing insights from more than thirty studies, the review sought to identify recurring patterns and themes relevant to LLM interactions with marginalized communities. Although the review is qualitative in nature, the search and screening procedures were conducted systematically to ensure transparency and reproducibility. Studies were identified through structured keyword searches across multiple academic databases and screened through sequential stages of title review, abstract evaluation, and full-text assessment.

Inclusion and Exclusion Criteria Data

Studies included in the review were primarily written in English. Research in other languages was excluded due to practical limitations related to translation and accessibility. Selected studies addressed at least one of the following topics: LLM interactions with LGBTQ or TGNB users; linguistic phenomena within LGBTQ communities; or Filipino LGBTQ linguistic practices such as Swardspeak. Both qualitative and quantitative studies were included. Quantitative research often measured performance outcomes such as pronoun accuracy or bias detection rates (Hossain et al., 2023; Tang et al., 2025). Qualitative research examined how LLMs discuss LGBTQ identities or respond to socially sensitive prompts (Das & Drolet, 2024; Scheuerman et al., 2025). Including both methodological approaches allowed

the review to capture a broader range of insights. Studies published between 2019 and early 2026 were considered. This time range was selected because LLM technology has developed rapidly in recent years, and earlier studies may reflect substantially different model capabilities.

Relevant studies were identified through academic databases, including Google Scholar, Scopus, and Web of Science. Search keywords included combinations of terms such as "LLM," "AI," "ChatGPT," "transgender," "non-binary," "LGBTQ," "linguistics," and "Filipino." These terms were combined in various ways to identify studies addressing both technological and sociolinguistic aspects of the research topic.

The search process was conducted in two phases. The first occurred during the final quarter of 2025, and the second was conducted in January 2026 to capture recently published studies. This approach helped ensure that emerging research themes were included. Studies were screened through a multi-stage process involving title review, abstract screening, and full-text evaluation. Both researchers independently reviewed candidate articles and discussed any disagreements until a consensus was reached. Because the study focused on qualitative synthesis rather than quantitative meta-analysis, statistical measures of inter-rater reliability were not applied.

Analysis and Coding

A research matrix was constructed using Microsoft Excel to organize key information from each study. The matrix included the following categories: author and year of publication; study focus; participants; LLM models used; methodological approach; key findings; and relevance to the conceptual framework. An inductive coding approach was used during analysis. Rather than beginning with predetermined categories, codes were developed based on recurring themes identified across the literature. During the coding process, passages from the reviewed studies were compared and grouped according to shared patterns.

Examples of resulting codes included "Western-centric bias," "Asian culture specific," and "LLM harm reproduction." These codes were then organized into broader thematic categories. Three major themes emerged from this analysis: first, the LLM limitations when addressing LGBTQ identities, second, the Linguistic practices within LGBTQ communities, and third, Filipino-specific linguistic and cultural phenomena. These themes guided the synthesis presented in the findings section.

Conceptual Framework

The following Intersectionality-based meta-synthesis framework was developed to examine the potential harms, limitations, and lack of fitness LLMs have when dealing with the Filipino TGNB. Intersectionality was used as the foundation because Ulnicane (2024) and Homan et al. (2024) found that it was useful for identifying LLMs' social biases and the resulting consequences. This is particularly so for the demographics in the intersections, as they tend to suffer more serious and unique harms and limitations. An example is LLMs erasing or missing specific cultural context around non-Western LGBTQ users (Hall et al., 2025; Ungless et al., 2025).

The Venn diagram illustrates the overlapping factors and resulting issues that make LLMs particularly unsafe, limited, and/or challenging for the Filipino TGNB. These are Non-Western Cultural Biases, Linguistic Limitations, and LGBTQ/TGNB social biases, which intersect to create compounded risks and harms like low-resource language issues (such as Filipino), native context gaps for the Filipino LGBTQ experience, and poor quality of service, like the potential difficulties LLMs have for understanding *Swardspeak*. A Thai user will find it difficult to use their native language with LLMs; an American lesbian will likely find herself falsely flagged for harm by using reclaimed slurs; a Filipino gay man will struggle to get an LLM to understand the connotations around *bakla*; and finally, in the center of the intersections, Filipino TGNB experience all these harms, plus more. This highlights the need for culturally sensitive, linguistically inclusive, and socially-bias-aware LLMs, and this framing of the sources/problems as a series of complex, overlapping factors affecting the Filipino TGNB uniquely was a primary guide for the researchers' coding and analysis of the reviewed studies.

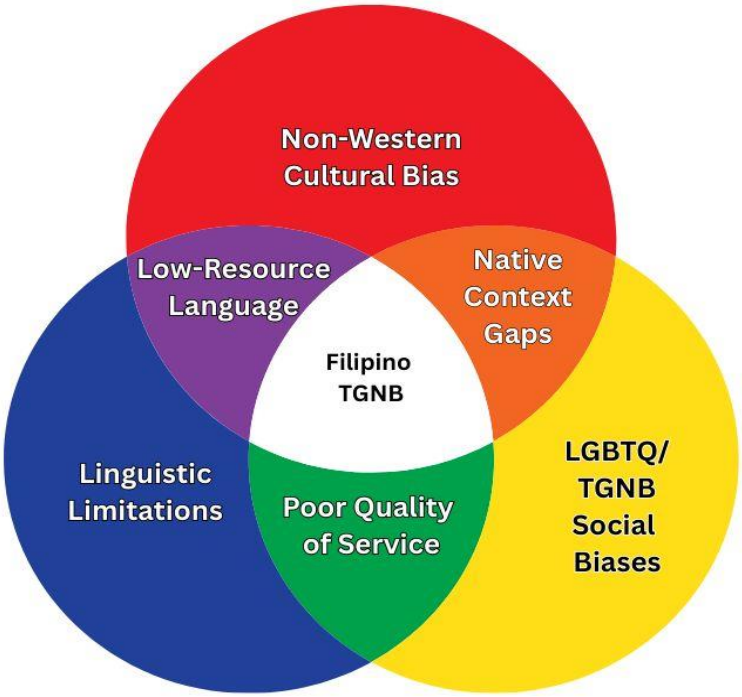

Figure 1. A Venn diagram of the Intersectionality-based framework.

Results

Limitations and Harms Toward TGNB Users

The reviewed literature consistently identified misgendering as one of the most common issues affecting TGNB interactions with LLM systems. Models sometimes replaced gender-neutral pronouns with binary alternatives or incorrectly inferred gender based on contextual cues (Hossain et al., 2023; Ovalle et al., 2023; Tang et al., 2025). Another recurring pattern involved the narrowing of TGNB narratives to a limited set of topics. Several studies reported that LLM outputs frequently focused on themes such as coming out, medical transition, or social

acceptance, even when prompts addressed unrelated topics (Ghosal et al., 2025). This pattern may limit the diversity of narratives presented about TGNB individuals. The literature also documented instances in which LLMs reproduced stereotypes or harmful language embedded within training data. In some cases, these patterns occurred even when prompts did not explicitly mention LGBTQ identities (Gupta et al., 2025).

TGNB Identities in Non-Western Contexts

Research addressing LGBTQ interactions with LLMs outside Western contexts remains limited, but existing studies suggest that cultural differences can influence model behavior. For example, Hall et al. (2025) found that Taiwanese LGBTQ participants emphasized the importance of family relationships in ways that were not always reflected in model responses. Within the Philippine context, cultural and linguistic factors may present additional challenges. Filipino language structures and identity categories do not always align directly with Western gender terminology. As a result, models trained primarily on English-language datasets may struggle to interpret locally specific expressions or cultural references.

Intersectionality as an Analytical Framework

Applying an intersectional perspective highlights how these linguistic, cultural, and social factors may interact. Rather than viewing each limitation separately, the framework suggests that combined factors can produce more complex forms of misalignment. For Filipino TGNB users, these intersecting dynamics may influence how effectively LLM systems interpret identity labels, linguistic expressions, and cultural contexts.

Discussions

The findings of this study highlight several recurring limitations in how LLM systems interact with TGNB identities and linguistic practices. When considered alongside the literature reviewed earlier, these findings suggest that many of the observed issues are not isolated technical problems but rather reflect broader structural patterns in how language models are trained and deployed. In particular, the results reinforce previous research indicating that LLMs often struggle with gender diversity, cultural context, and non-standard linguistic forms. When these factors intersect, as in the case of Filipino TGNB users, the resulting challenges can become more complex.

One of the most consistently documented limitations involves the handling of gender-neutral pronouns and non-binary identities. As noted in the findings, several studies reported instances of misgendering, where LLMs incorrectly replace gender-neutral pronouns such as “they” with binary pronouns like “he” or “she” (Hossain et al., 2023; Ovalle et al., 2023; Tang et al., 2025). These findings are consistent with earlier literature suggesting that training datasets often contain far fewer examples of non-binary pronouns compared with binary gender references. Because LLMs rely on probabilistic associations learned from training data, these imbalances may lead models to default to more common linguistic patterns. The persistence of such issues, even in more recent models, indicates that improvements in model architecture alone may not fully resolve representational disparities within training corpora.

Beyond pronoun usage, the literature also points to limitations in how LLMs construct narratives about TGNB individuals. As discussed in the literature review, Ghosal et al. (2025)

described the phenomenon of *Trans Broken Arm Syndrome*, in which model responses repeatedly frame TGNB identities primarily through the lens of transition, medical procedures, or social acceptance. The findings of the present review reinforce this observation. Across several studies, LLM outputs tended to emphasize a relatively narrow set of themes when discussing transgender or non-binary individuals. This pattern suggests that the narratives most visible within training datasets may disproportionately influence how models generate content about marginalized identities. Consequently, LLM-generated discourse may unintentionally reproduce limited or stereotypical portrayals of TGNB experiences.

Another important theme emerging from the findings concerns the influence of training data on the reproduction of social biases. Prior research has emphasized that LLM outputs often reflect patterns embedded within their training sources (Gupta et al., 2025; Tomasev et al., 2021; Zhou, 2024). Because large-scale datasets are frequently drawn from internet text, they inevitably contain both supportive and discriminatory language. The literature reviewed in this study indicates that LLMs may reproduce these patterns in subtle ways, even when explicit safeguards are implemented. For example, Bergstrand & Gambäck (2024) observed that models trained on Norwegian data sometimes reproduced patterns of discrimination present in that linguistic environment. Similarly, Gamboa & Lee (2025) found that multilingual models interacting with Philippine prompts could reproduce locally embedded homophobic stereotypes. These findings suggest that bias mitigation strategies must address not only overtly harmful language but also the broader social context represented in training data.

Cultural context represents another key dimension shaping LLM performance. Many existing models are trained primarily on English-language text originating from Western contexts, which may limit their ability to interpret culturally specific norms or social dynamics in other regions. Hall et al. (2025), for instance, demonstrated that Taiwanese LGBTQ users often emphasize family relationships and social obligations in ways that differ from Western narratives of individual identity. When LLMs fail to recognize these cultural nuances, their responses may appear misaligned with users' lived experiences. The findings of this study suggest that similar challenges may arise in the Philippine context, where local linguistic and cultural dynamics shape how LGBTQ identities are expressed.

The Filipino linguistic environment presents several unique challenges for language models. Unlike English, Filipino languages do not always contain widely recognized equivalents for certain gender identity categories such as "non-binary" or "transgender" (Pitargue, 2021). Instead, historically rooted terms such as "*bakla*" may combine aspects of gender expression and sexual orientation (Botero, 2022). Because these linguistic categories differ from Western terminology, LLM systems trained primarily on English-language frameworks may struggle to interpret Filipino LGBTQ discourse accurately. This misalignment can lead to misunderstandings when models attempt to map local identity expressions onto Western conceptual categories. The presence of *Swardspeak* within Filipino LGBTQ communities further illustrates the importance of linguistic context. As described in the literature review, *Swardspeak* is characterized by creative word formation processes, code-switching, and frequent references to pop culture (Nuncio et al., 2021; Ulla et al., 2024). These linguistic features allow speakers to signal community membership and express identity in socially meaningful ways. However, because *Swardspeak* relies heavily on contextual

knowledge and rapidly evolving vocabulary, it may pose particular challenges for language models trained on standardized written text. If LLM systems cannot accurately interpret such expressions, Filipino LGBTQ users may encounter limitations when attempting to communicate nuanced ideas through AI interfaces.

The intersectionality-based framework proposed in this study provides one way to conceptualize how these challenges interact. Rather than viewing linguistic limitations, cultural context gaps, and LGBTQ-related biases as separate issues, the framework highlights how they may combine to produce distinct forms of misalignment. For example, a Filipino TGNB user interacting with an LLM may encounter difficulties related to pronoun recognition, misunderstandings of *Swardspeak* expressions, and culturally inappropriate interpretations of gender identity. Each of these issues alone may appear manageable, but their combined effects can significantly shape user experiences with AI systems.

Existing research on AI fairness increasingly emphasizes the importance of intersectional perspectives when evaluating technological systems. Homan et al. (2024) demonstrated that incorporating diverse perspectives among safety raters can improve the detection of problematic model outputs. Similarly, Ulnicane (2024) argued that discussions of AI governance must consider social inequalities beyond economic or technical dimensions. The framework developed in the present study extends these insights by applying an intersectional lens to the specific context of Filipino TGNB users interacting with LLM technologies. Importantly, the findings of this study do not suggest that LLMs are inherently incapable of interacting safely with marginalized communities. Rather, they indicate that current systems may require additional cultural and linguistic sensitivity to operate effectively across diverse user populations. As LLM technologies continue to expand globally, addressing these challenges will likely require collaboration between AI developers, linguists, and scholars specializing in gender and sexuality studies.

Overall, the discussion underscores the importance of recognizing how technological systems intersect with social and cultural realities. The experiences of Filipino TGNB users highlight broader questions about whose language, identities, and cultural contexts are represented within AI systems. Addressing these questions will be essential for ensuring that future language technologies support inclusive and equitable communication environments.

Conclusion

This study synthesized existing literature examining LLM interactions with LGBTQ and TGNB identities while considering linguistic and cultural factors relevant to the Filipino context. The findings suggest that LLM systems may encounter difficulties related to pronoun usage, narrative framing, and culturally specific linguistic practices. By proposing an intersectionality-based framework, the study highlights how linguistic limitations, cultural context gaps, and social biases may interact in shaping these outcomes. Addressing these issues will likely require continued interdisciplinary collaboration among AI researchers, linguists, and LGBTQ scholars to ensure that emerging language technologies can serve diverse communities more equitably.

The study highlights several implications for AI development and research. First, greater attention may be needed to ensure that datasets include diverse linguistic and cultural contexts. Second, involving LGBTQ participants directly in evaluation processes may help identify forms of bias that are not immediately visible in automated testing. This review focused primarily on English-language research, which may have excluded studies conducted in other languages. Additionally, empirical studies involving Filipino TGNB participants remain limited. Future research could explore how multilingual LLMs perform when interacting directly with Filipino LGBTQ users or when responding to prompts written in *Swardspeak*.

References

Abulog, M., Chung, K. A. C., Manuel, C. S., Manuel, L. D. M., & Roca, F. C. A. (2023). A Research on the Usage and Deviation to Gay Lingo of Different Professionals Belonging to the LGBTQ++ in NCR. *International Journal of Latest Research in Humanities and Social Science (IJLRHSS), 06*(08), 150–154.

Almoite, A. D. (2025). Unfolding a Unique Tongue: A Morphological Formation of Swardspeak in a State University. *K@ta*, *27*(1), 34–47. https://doi.org/10.9744/kata.27.1.34-47

Anderson, S. M. (2020). Gender Matters: The Perceived Role of Gender Expression in Discrimination Against Cisgender and Transgender LGBQ Individuals. *Psychology of Women Quarterly, 44*(3), 323–341. https://doi.org/10.1177/0361684320929354

Atienza, P. M. L. (2023). "I Look at How They Write Their Bio and I Judge from There": Language and Class Among Middle-Class Queer Filipino Digital Socialities in Manila. *International Journal of Communication*, *17*, 2498–2513.

Bergstrand, S., & Gambäck, B. (2024). Detecting and Mitigating LGBTQIA+ Bias in Large Norwegian Language Models. *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing* (GeBNLP), 351–364. https://doi.org/10.18653/v1/2024.gebnlp-1.22

Borba, R. (2019). The interactional making of a "true transsexual": Language and (dis)identification in trans-specific healthcare. *International Journal of the Sociology of Language, 2019*(256), 21–55. https://doi.org/10.1515/ijsl-2018-2011

Botero, R. (2022). Genderless: What it means to be non-binary in the Philippines and understanding SOGIE. *Philstar Life*. https://philstarlife.com/self/740382-genderless-filipino-non-binary-experience?

Brown, C. (2022). *The Politics of Community Language Change: A Computational Analysis of Language Norms in an Online Trans Community*. University of California.

Brown, C., Zimman, L., & Todd, S. (2025). Dynamics of Language Change: A Mixed-Methods Analysis of Language in an Online Transgender Community. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*, 289–306.

Ciston, S. (2019). Intersectional AI Is Essential. *Journal of Science and Technology of the Arts, 11*, 3–8. https://doi.org/10.7559/citarj.v11i2.665

Das, R. K., & Drolet, B. C. (2024). Assessment of Artificial Intelligence Chatbot Attitudes Toward LGBTQ+ Individuals. *Journal of Adolescent Health, 74*(6), 1264–1266. https://doi.org/10.1016/j.jadohealth.2024.02.030

de Carvalho Souza, M. E., & Weigang, L. (2025). Grok, Gemini, ChatGPT and DeepSeek: Comparison and applications in conversational artificial intelligence. *Inteligencia Artificial, 2*(1). https://doi.org/10.5281/zenodo.14885243

Dhingra, H., Jayashanker, P., Moghe, S., & Strubell, E. (2023). Queer People are People First: Deconstructing Sexual Identity Stereotypes in Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2307.00101

Dorn, R., Kezar, L., Morstatter, F., & Lerman, K. (2024). Harmful Speech Detection by Language Models Exhibits Gender-Queer Dialect Bias. *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–12. https://doi.org/10.1145/3689904.3694704

Edwards, J., Clark, L., & Perrone, A. (2021). LGBTQ-AI? Exploring Expressions of Gender and Sexual Orientation in Chatbots. *CUI 2021 - 3rd Conference on Conversational User Interfaces*, 1–4. https://doi.org/10.1145/3469595.3469597

Gamboa, L. C. L., & Lee, M. (2025). Filipino Benchmarks for Measuring Sexist and Homophobic Bias in Multilingual Language Models from Southeast Asia. *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)*, 123–134.

Ghosal, A., Gupta, A., & Srikumar, V. (2025). *Unequal Voices: How LLMs Construct Constrained Queer Narratives*. Arxiv. https://doi.org/10.48550/arXiv.2507.15585

Goethals, S., Rhue, L., & Sundararajan, A. (2026). Fairness principles across contexts: evaluating gender disparities of facts and opinions in large language models. *AI and Ethics*, *6*(1), 41. https://doi.org/10.1007/s43681-025-00876-5

Gupta, O., Marrone, S., Gargiulo, F., Jaiswal, R., & Marassi, L. (2025). Understanding Social Biases in Large Language Models. *AI*, *6*(5), 106. https://doi.org/10.3390/ai6050106

Hall, A., Wang, C.-L., & Tseng, Y.-C. (2025). *Culturally Adaptive Chatbot Design for LGBTQ+ Support in Family-Centric Societies*. SSRN. https://doi.org/10.2139/ssrn.5574177

Haxvig, H. A., D'Andrea, V., & Teli, M. (2025). Synthetic Dreams in Barbie Land: Speculative Queer Adventures with Feminist LLM-Generated Personas. *Companion Publication of the 2025 ACM Designing Interactive Systems Conference*, 379–385. https://doi.org/10.1145/3715668.3736361

Homan, C., Serapio-Garcia, G., Aroyo, L., Diaz, M., Parrish, A., Prabhakaran, V., Taylor, A., & Wang, D. (2024). Intersectionality in AI Safety: Using Multilevel Models to Understand Diverse Perceptions of Safety in Conversational AI. *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)*, 131–141. https://aclanthology.org/2024.nlperspectives-1.15/

Hossain, T., Dev, S., & Singh, S. (2023). MISGENDERED: Limits of Large Language Models in Understanding Pronouns. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5352–5367. https://doi.org/10.18653/v1/2023.acl-long.293

Jacobsen, K., Devor, A., & Hodge, E. (2022). Who Counts as Trans? A Critical Discourse Analysis of Trans Tumblr Posts. *Journal of Communication Inquiry*, *46*(1), 60–81. https://doi.org/10.1177/01968599211040835

Khan, A. J., & Anwar, D. N. (2024). Identity Denial in The Stylistic Features of Transgender Language. *International Journal of Contemporary Issues in Social Sciences*, *3*(2). Retrieved from https://ijciss.org/index.php/ijciss/article/view/736

Nuncio, R. V, Pamittan, G. B., Corpuz, D. R., & Ortinez, E. V. (2021). Jokla and Jugels: A Comparative Analysis of the Construction of Popular and Hiligaynon Gay Words. *Humanities Diliman*, *18*(2), 37–64.

Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. *Scientific Reports, 15*(1), 13755. https://doi.org/10.1038/s41598-025-98483-1

Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K.-W., Galstyan, A., Zemel, R., & Gupta, R. (2023). "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. *2023 ACM Conference on Fairness Accountability and Transparency*, 1246–1266. https://doi.org/10.1145/3593013.3594078

Pitargue, A. (2021). How non-binary Filipinos reconcile their identities with their language's lack of LGBT terms. *CBC/Radio-Canada*. https://www.cbc.ca/news/canada/british-columbia/filipino-nonbinary-tagalog-language-1.6119416

Scheuerman, M. K., Weathington, K., Petterson, A., Doyle, D. T., Das, D., DeVito, M. A., & Brubaker, J. R. (2025). Transphobia Is in the Eye of the Prompter: Trans-Centered Perspectives on Large Language Models. *ACM Transactions on Computer-Human Interaction*, *32*(5), 1–42. https://doi.org/10.1145/3743676

Tang, X., Ding, Y., Yang, Z., Chen, Y., Gu, Y., Yang, W., Ju, M., Cao, X., Liu, Y., & Zhang, W. (2025). Do They Understand Them? An Updated Evaluation on Nonbinary Pronoun Handling in Large Language Models. *arXiv.* https://doi.org/10.48550/arXiv.2508.00788

Tomasev, N., McKee, K. R., Kay, J., & Mohamed, S. (2021). Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 254–265. https://doi.org/10.1145/3461702.3462540

Ulla, M. B., Macaraeg, J. M., & Ferrera, R. E. (2024). 'What's the word? That's the word!': linguistic features of Filipino queer language. *Cogent Arts & Humanities*, *11*(1), 2322232. https://doi.org/10.1080/23311983.2024.2322232

Ulnicane, I. (2024). Intersectionality in Artificial Intelligence: Framing Concerns and Recommendations for Action. *Social Inclusion*, *12*, 7543. https://doi.org/10.17645/si.7543

Ungless, E. L., Dev, S., Bennett, C. L., Gulotta, R., Bastings, J., & Denton, R. (2025). Amplifying Trans and Nonbinary Voices: A Community-Centred Harm Taxonomy for LLMs. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 20503–20535.

West, A., Wada, K., & Strong, T. (2021). Authenticating and Legitimizing Transgender and Gender Non-conforming Identities Online: A Discourse Analysis. *Journal of LGBT Issues in Counseling*, *15*(2), 195–223. https://doi.org/10.1080/15538605.2021.1914275

Wilson, H., Malik, A., & Thompson, S. (2024). How Transgender Adolescents Experience Expressing Their Gender Identity Around New People: An Interpretative Phenomenological Analysis. *Journal of Adolescent Research*, *39*(1), 30–52. https://doi.org/10.1177/07435584211043879

Zhou, A. (2024). Queer Bias in Natural Language Processing: Towards More Expansive Frameworks of Gender and Sexuality in NLP Bias Research. *AI in Education, Culture, Finance, and War*, *2*. https://doi.org/10.60690/p45ma787

Zimman, L. (2019). Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, *2019*(256), 147–175. https://doi.org/10.1515/ijsl-2018-2016

Zimman, L., & Hayworth, W. (2020). How we got here: Short-scale change in identity labels for trans, cis, and non-binary people in the 2000s. *Proceedings of the Linguistic Society of America*, *5*(1), 499. https://doi.org/10.3765/plsa.v5i1.4728