Guardrails, not Guidance: Understanding Responses to LGBTQ+ Language in Large Language Models

Joshua Tint
Arizona State University
jrtint@asu.edu

Abstract

Language models have integrated themselves into many aspects of digital life, shaping everything from social media to translation. This paper investigates how large language models (LLMs) respond to LGBTQ+ slang and heteronormative language. Through two experiments, the study assesses the emotional content and the impact of queer slang on responses from models including GPT-3.5, GPT-4o, Llama2, Llama3, Gemma and Mistral. The findings reveal that heteronormative prompts can trigger safety mechanisms, leading to neutral or corrective responses, while LGBTQ+ slang elicits more negative emotions. These insights punctuate the need to provide equitable outcomes for minority slangs and argots, in addition to eliminating explicit bigotry from language models.

1 Introduction

Bias in language reflects and reinforces social norms, shaping perceptions of identity and inclusivity in both human- and machine-mediated communication. As large language models increasingly mediate our conversations, the biases encoded in these systems gain the power to construct and perpetuate inequities (Felkner et al., 2023; Ungless et al., 2023). Queer communities, in particular, are heavily impacted by biased language technologies. Online spaces often serve as vital forums for connection, support, and expression for LGBTQ+ people, disproportionately exposing them to any potential LLM bias (Leap, 2023). This paper examines biases in the responses of language technologies to two distinct kinds of linguistic expression related to the queer community: heteronormative language, and queer slang. Understanding these biases is essential to ensure that these systems support fairness and inclusion, rather than amplifying existing inequities.

Recent research has highlighted the pervasive biases encoded in LLMs. This includes reinforcing harmful stereotypes, such as associating particular occupations with particular genders or disproportionately flagging minority dialects as toxic (Zhao et al., 2019; Sap et al., 2019). Benchmarks like WinoQueer have shed light on anti-queer biases in model outputs, calling for community-driven evaluations to improve fairness (Felkner et al., 2023). Although efforts to mitigate bias have focused on safety mechanisms and debiasing techniques, these approaches primarily address overt discrimination and fail to account for subtler forms of bias, such as those found in responses to non-standard linguistic features like queer slang (Lin et al., 2024).

This work builds on prior research in two clear ways. Primarily, we focus on prompts containing language used by queer people, rather than queer topics or scenarios explicitly involving queer themes. Additionally, we measure the emotional content of model responses, providing a more nuanced view into implicit bias shown by language models. Together, this approach allows us to move beyond surface-level evaluations of fairness by examining how language models react to subtle linguistic markers of identity.

In particular, this paper addresses gaps in the current research by focusing on two central questions:

- **RQ1:** How does the emotional content of LLM-generated responses vary when prompted with heteronormative versus non-heteronormative language?
- **RQ2:** How does the presence of LGBTQ+ slang in prompts influence the emotional content of LLM outputs?

The findings presented here reveal critical gaps in current fairness approaches. While safety mechanisms neutralize bias in responding to overt heteronormative prompts, they fail to address systemic biases in responses to queer slang, which often elicit disproportionately negative emotional labels. These results highlight the limitations of existing debiasing efforts and underscore the importance of improving LLM outputs for language used by marginalized communities.

To foster truly inclusive NLP systems, future research and development must prioritize the equitable representation of minority linguistic forms. By expanding evaluation frameworks to account for nuanced biases, we can ensure that LLMs reflect the diversity of human language and support marginalized voices in digital spaces.

The primary contributions of this work are:

- We introduce a unique embedding-based clustering approach, using Mahalanobis distance, to

Proceedings of the Queer in AI Workshop, pages 6–16
May 4, 2025 ©2025 Association for Computational Linguistics

quantify the presence and influence of LGBTQ+ slang in prompts.

- Through emotional classification of LLM-generated outputs, we reveal that queer slang prompts elicit disproportionately negative emotional labels compared to heteronormative or neutral language.

- We provide evidence that current safety mechanisms in LLMs fail to address deeper systemic biases, highlighting the limitations of existing approaches in achieving inclusivity for marginalized linguistic communities.

2 Related Work

2.1 Queer Slang

Queer communities have developed a rich linguistic tradition, characterized by unique syntax, grammar, and slang, often distinct from cisgender and heterosexual norms. Historical examples include Polari, a coded language used by LGBTQ+ individuals when their identities were criminalized, elements of which persist in modern slang such as "zhush" and "camp" (Baker, 2003). The advent of digital communication has expanded the reach of queer slang, enabling phrases such as "spill the tea" and "throw shade" to gain mainstream recognition (Karabayik and Saavedra, 2022). However, queer slang is underrepresented in large language model training corpora due to its rapid evolution, niche contexts, and prevalence in semi-private spaces (Ungless et al., 2023). Additionally, queer slang intersects with African American Vernacular English (AAVE) in phrases like "queen" and "chile," complicating biases due to the overlapping marginalization of these dialects (Leap, 2023; Blackburn, 2005).

2.1.1 Heteronormativity in Language

Heteronormativity is a broad phenomenon which encompasses assumptions of heterosexuality, traditional gender roles, and binary gender norms. In language, heteronormativity can reveal itself in a variety of ways. Primarily, heteronormative language encodes normative sexual and gender behaviors. Marchia and Sommer provide a taxonomy of heteronormativity which includes many distinct forms (Marchia and Sommer, 2019). 1 contains examples of heteronormative language, including their categorization within Marchia and Sommer's framework. The four categories presented by Marchia and Sommer are gendered heteronormativity, or the assumption of gender roles, cisnormative heteronormativity, or the assumption of cisgenderism as the default, heterosexist heteronormativity, which is the assumption of heterosexuality as the default, and hegemonic heteronormativity, which encompasses any other kind of cultural sphere which leads to other kinds of heteronormativity. Addressing such biases is critical for creating inclusive NLP systems capable of understanding and generating non-normative expressions.

Vasquez operates within this framework to offer a simplified and unified definition, offering that heteronormative speech is that which " creates boundaries of normative sexual behavior, and relate to behaviors and feelings against violations of these norms" (Vásquez et al., 2022). This categorization is useful because it allows for a clear, binarized, "heteronormative-or-not" classification, and thus will form the basis of heteronormativity in this work, although the work presented by Marchia and Sommer helps to understand the taxonomy of heteronormativity.

2.2 Bias in LLMs

Biases in LLMs arise during data collection, model development, and evaluation (Dai et al., 2024). Gender bias, for example, persists even in advanced models like BERT, as shown by associations linking professions to gender stereotypes (Bolukbasi et al., 2016; Zhao et al., 2019). Similarly, racial bias has been identified in sentiment analysis systems, where names or references associated with marginalized groups receive disproportionately negative sentiment (Kiritchenko and Mohammad, 2018). Dialects such as AAVE are often over-moderated in content moderation tasks, further marginalizing non-standard speech patterns (Sap et al., 2019).

3 Method

3.1 Experiment 1

Experiment 1 focuses on understanding the tone and emotional response of language models to heteronormative versus non-heteronormative prompts, answering RQ1: "How does the presence of LGBTQ+ slang in prompts influence the emotional content of LLM-generated outputs?" The models used were: GPT-3.5, GPT-4o, Llama2, Llama3.2, Gemma, Gemma2, and Mistral. These models were selected to represent a large contingent of LLM families, and to represent a diverse array of parameter sizes. Two different experiments shared similar methodologies: each model was prompted with text emulating a user input, exhibiting a varying degree of heteronormativity. However, the experiments differed in how the models were prompted.

3.1.1 Experiment 1.1

In order to obtain high-quality prompts, we used Vasquez's HeteroCorpus, which is a dataset of 7,266 posts from X.com (formerly Twitter) tagged for heteronormativity, which they represent as a boolean variable: "heteronormative" or "non-heteronormative" (Vásquez et al., 2022). From this corpus, we pulled a random sample of 500 posts, including 250 heteronormative posts and 250 non-heteronormative posts, and then fed each of these posts to the language models in the experiment, recording their responses.

Example	Explanation
If a doctor has recently graduated medical school, then he can expect a lower salary.	This exhibits gendered heteronormativity. While "doctor" is gender-neutral, the use of "he" presupposes that the doctor is male.
When a woman gets married, she will want her husband to be kind.	This exhibits heterosexist heteronormativity, by implying that a woman must have a husband.
Does he have a husband or wife?	This avoids heterosexist heteronormativity by acknowledging that a man could have a husband. However, it exhibits cisnormative heteronormativity by reinforcing a gender binary with the phrase "husband or wife."

Table 1: Examples of heteronormative language and their classifications.

To gauge the emotional content of each response, an emotional classifier was trained using RoBERTa-Base on Google's GoEmotions dataset. RoBERTa-Base was chosen because it is a transformer-based model known for its strength in text classification tasks, especially in tasks that involve nuanced language that could contain multiple sentiments or subtle emotional undertones (Tan et al., 2022; Liu, 2019). Since RoBERTa-Base is pre-trained on large-scale general language corpora and has demonstrated high performance across NLP benchmarks, it provides a strong foundation for accurately detecting emotional cues in language. Additionally, RoBERTa-Base's architecture is specifically suited for tasks requiring high sensitivity to context, a critical feature when analyzing emotionally rich content such as social media posts (Petroni et al., 2020). The GoEmotions dataset, used to fine-tune the RoBERTa-Base model, offers 28 fine-grained emotional labels, allowing the classifier to identify a wide array of emotional responses. The dataset itself consists of social media posts, making it an ideal match for the content in HeteroCorpus, as both contain similar linguistic styles and content structures. This classifier outputs confidence values for each of the 28 emotional categories, providing a nuanced view of emotional content and facilitating a more detailed analysis of the relationship between language heteronormativity and model emotional responses. The trained RoBERTa model was able to achieve $> 94\%$ accuracy for each emotion label except for neutral, which was labeled at a 74% accuracy.

3.1.2 Experiment 1.2

In this experiment, prompts were sourced from the Quora question pairs dataset (Chen et al., 2017). This corpus contains over 400,000 questions across a variety of topics, in a paired format, annotated as to whether the questions are equivalent or not. "Equivalent," in this case, refers to whether the questions paraphrase one another. Paired questions are advantageous, because they control for a variety of factors, such as topic, which might confound the outcome of a non-paired experiment such as Experiment 1.1. In addition, they allow for direct comparison between identical questions that primarily differ only in their use of heteronormative language. Unfortunately, the Quora dataset is not manually tagged by heteronormativity, so in order to find identical questions which had differing heteronormativity, an automated system had to be built and deployed. The question pairs were first filtered if they were tagged as equivalent and if one or both contained a set of potentially heteronormative keywords. This was a list of gendered terms like "policewoman" and "mankind," along with equivalents designed to be specifically non-heteronormative, such as "partner" (as opposed to "boyfriend" or "girlfriend) and "congressperson" (rather than "congressman"). Following this step, an automated system was used to determine the relative heteronormativity of each question. GPT-4o fed the Vasquez definition of heteronormativity along with three annotated examples (Prompt 1). It was then prompted with both questions and asked to decide whether one was more heteronormative than the other, or whether they were equivalently heteronormative. Consistent prompting with Vasquez's definitions was included to improve alignment with HeteroCorpus and response quality. Ultimately, 1398 equivalent question pairs with differing heteronormativity were extracted.

Following this, responses were collected and evaluated similarly to Experiment 1.1. For each prompt question, a response was collected for each of the LLMs in the experiment. The GoEmotions-trained model was used to give emotional classifications for each label on each response. The difference between each emotion confidence value for the response to the heteronormative prompt against the response to the non-heteronormative prompt was calculated to get a paired value.

3.2 Experiment 2

The primary goal of Experiment 2 is to answer RQ2: "How does the presence of LGBTQ+ slang in prompts influence the emotional content of LLM-generated outputs?" In contrast to experiment 1, here we examine the presence of LGBTQ+ slang rather than the absence of heteronormativity. This builds on the results from experiment 1 by examining a broader range of linguistic features, but due to a lack of high-quality hand-tagged data on LGBTQ+ slang, this relies on a more general approach.

We propose a method for evaluating LGBTQ+ slang through embeddings. We begin with a list of 100 base LGBTQ+ slang terms. These terms are collated from a variety of sources that identify queer slang (Cantina, 2020; Jacobs, 1997; Vecchio, 2021; Laing, 2021; Kulick, 2000; Morgan, 2017; Simes, 2005; Rosales and Careterro, 2019). Terms with common alternate interpretations which eclipse their LGBTQ+ interpretations, such as "read" or "queen," were filtered out. In total, 57 terms were collected (In Appendix 8.2). The embeddings of all of these terms was measured from the popular `all-MiniLM-L6-v2` transformer. This creates a cluster of embeddings representing LGBTQ+ slang. From the set of the embeddings of LGBTQ+ slang terms, we define the function $F(t)$ which gives the Mahalanobis distance from the embedding of the text t to the LGBTQ+ slang embedding cluster. While Mahalanobis distance can be sensitive to outliers, it's well-suited for measuring the relative closeness of terms within the LGBTQ+ slang embedding cluster due to its ability to account for feature variance, ensuring that both common and niche slang expressions are represented. Additionally, it is better able to identify and account for the "shape" of a cluster of embeddings, making it well suited for point-to-cluster comparisons.

One potential issue with this method is that it will not reflect any syntactic features to LGBTQ+ slang, only the semantic and lexical ones that are incorporated into the single-word embeddings. Syntax is a known feature of LGBTQ+ slang, though it is usually not exhibited exclusively without the presence of other features. The list of LGBTQ+ slang terms is also by no means exclusive, and is meant to capture a broad cross-section of English slang terms which may have been used in LLM training data. However, by measuring embeddings, even LGBTQ+ slang terms not present on the base list can be measured as similar to the cluster.

We then select a random sample of 500 question pairs from the Quora paired question dateset. Because LGBTQ+ slang and not heteronormativity is the focus of this experiment, we employed no filtering measures such as in Experiment 1.2. We measure the F-score of each question in the sample. We record the result of each question in this dataset for each of the LLMs in the experiment. We then use the GoEmotions emotional classifier to measure the sentiment of each response, similar to Experiment 2. Rather than grouping responses by prompt heteronormativity, however, in this experiment we measure the correlation of different kinds of emotions to F-score.

4 Results

4.1 Experiment 1

4.1.1 Experiment 1.1

In order to measure the effects that prompt heteronormativity had on emotional content, we calculated the difference of means effect size of each emotional score for each model between the average emotion confidence score when given a heteronormative prompt against the average emotion confidence score when given a non-heteronormative prompt. The effect size was a standard Cohen's d.

Because many of these labels are similar or fine-grained, in order to get a broader picture on the results, we produced two more scores, "positive" and "negative" which were sums of other individual emotions‘ scores. "Disapproval," "annoyance," "nervousness," "disappointment," "grief," "disgust," "sadness," "anger," and "remorse" were coded as negative whereas "joy," "gratitude," "excitement," "approval," "caring," "relief," "pride," "amusement," "love," and "admiration" were coded as positive.

Complete results for each emotional label effect size for each model can be seen in Table 2.

4.1.2 Experiment 1.2

For this experiment, we study the paired effect size, measured in standardized mean difference, between each heteronormative sample's emotion scores, and its equivalent non-heteronormative sample's emotion scores. Positive effect sizes indicate emotions which occurred more prevalently in heteronormative data, whereas negative effect sizes indicate emotions which occurred more prevalently in non-heteronormative data. The same "positive" and "negative" labels were used from the prior subexperiment. We also computed average effect size scores across all models for each emotional label, in order to examine overall trends for emotions.

Complete scores for each emotion label across each model can be found in Table 3.

4.2 Experiment 2

In order to measure the relationship between prompt F-score and the emotional content of LLM responses, we measure the $\Delta F = F(q_1) - F(q_2)$ of each question pair (q_1, q_2). We then measured the score $\texttt{Emotion}_e(r)$ which represents the score for the emotional label e of the response r given by the classifier. From that, we compute $\Delta\texttt{Emotion}_e = \texttt{Emotion}_e(r_1) - \texttt{Emotion}_e(r_2)$ of each response pair (r_1, r_2). In order to track the correlation between F scores and emotions, we simply calculate the proportion $\Delta\texttt{Emotion}/\Delta F$ for the responses to each question pair.

Similar to experiment 2, we also created the meta-labels "positive" and "negative," which had confidence scores equal to the summed confidence scores of the same labels as in the previous experiment. This again

Emotion	GPT-3.5	GPT-4o	Llama2	Llama3.2	Gemma	Gemma2	Mistral	Average
joy	-1.59	-0.77	0.40	-0.09	0.74	0.49	-2.23	-0.44
gratitude	-1.27	-3.72	-3.49	-4.88	0.03	-1.44	0.12	-2.09
excitement	-0.90	-0.63	0.36	-0.02	-0.46	-0.58	-3.61	-0.83
confusion	-0.60	0.87	-0.63	0.49	-0.46	0.07	0.35	0.01
approval	-0.41	-0.17	0.37	-0.45	-0.08	-0.25	-0.08	-0.15
optimism	-0.23	-0.90	0.32	-0.56	-0.05	0.17	-0.33	-0.23
disapproval	-0.23	-0.85	-1.46	0.01	1.48	-0.17	0.16	-0.15
caring	-0.21	-0.15	1.45	0.14	-0.35	-0.27	-1.05	-0.06
annoyance	-0.20	-0.38	-0.35	0.20	0.90	0.89	0.10	0.16
nervousness	-0.16	0.15	-0.02	0.08	-0.05	0.09	-0.04	0.01
relief	-0.15	-0.06	0.17	-0.06	0.27	0.25	-0.39	0.00
realization	-0.09	-0.07	0.03	-0.03	0.18	0.13	0.57	0.10
fear	-0.06	-0.11	0.30	-0.01	-0.06	-0.06	0.00	0.00
disappointment	-0.04	-0.12	-1.98	0.24	0.60	0.12	0.20	-0.14
desire	-0.04	-0.10	0.06	-0.27	-0.56	-0.41	0.26	-0.15
grief	-0.01	-0.12	-0.31	0.11	0.01	-0.16	-0.05	-0.08
disgust	0.01	0.03	-0.12	0.02	0.20	0.05	0.04	0.03
sadness	0.01	0.17	-2.99	-0.08	-0.00	0.27	-1.06	-0.53
anger	0.03	-0.21	-0.22	0.04	0.44	0.30	0.03	0.06
embarrassment	0.03	0.10	-0.26	0.04	0.12	0.02	0.09	0.02
pride	0.04	0.03	0.11	-0.06	0.11	0.06	-0.16	0.02
amusement	0.14	-0.02	0.41	0.29	0.02	0.02	0.15	0.14
remorse	0.20	0.35	-1.00	-0.17	-1.37	-0.54	-0.67	-0.46
love	0.23	-1.24	-0.12	0.42	-0.11	-0.32	-0.18	-0.19
curiosity	0.33	0.04	-0.51	-0.88	-1.24	-0.72	1.40	-0.23
neutral	0.47	1.12	0.27	0.63	-0.89	-0.32	0.66	0.28
surprise	0.57	0.42	0.01	0.04	-0.01	0.23	0.68	0.28
admiration	1.24	0.04	0.86	-0.55	0.56	0.42	0.79	0.48
NEGATIVE	-0.40	-0.98	-8.46	0.44	2.21	0.86	-1.29	-1.09
POSITIVE	-3.12	-7.60	0.84	-5.83	0.69	-1.45	-6.96	-3.35

Table 2: The *difference-of-means* effect size of heteronormativity on emotion scores. Negative figures are highlighted in red and indicate labels more associated with the non-heteronormative responses. Positive figures are highlighted in green and indicate labels more associated with the heteronormative responses.

Emotion	GPT-3.5	GPT-4o	Llama2	Llama3.2	Gemma	Gemma2	Mistral	Average
joy	-0.03	-0.26	-0.11	-0.38	-0.19	-0.23	-0.34	-0.22
gratitude	-0.02	-0.06	0.24	-0.29	0.05	0.02	-0.42	-0.07
excitement	-0.08	-0.05	-0.16	0.07	-0.27	-0.14	-0.40	-0.15
confusion	-0.10	-0.25	0.18	0.48	-0.28	0.14	0.05	0.03
approval	-0.27	-0.42	-0.33	-0.49	-0.29	-0.35	-0.36	-0.36
optimism	-0.30	-0.17	-0.41	0.16	0.16	0.03	0.01	-0.07
disapproval	0.26	-0.24	-0.05	-0.29	-0.52	-0.09	0.21	-0.10
caring	-0.37	0.25	-0.10	-0.33	0.28	0.00	0.37	0.02
annoyance	0.47	-0.21	-0.19	-0.39	0.31	0.27	0.34	0.09
nervousness	-0.46	0.12	-0.29	0.32	0.03	0.10	0.32	0.02
relief	-0.19	-0.23	-0.01	-0.41	-0.03	-0.10	0.30	-0.10
realization	-0.12	-0.20	-0.29	-0.29	-0.10	-0.11	0.17	-0.13
fear	-0.37	0.03	-0.26	0.34	-0.05	-0.29	0.09	-0.07
disappointment	0.28	-0.21	-0.38	0.30	0.32	0.09	0.32	0.10
desire	-0.41	0.05	-0.50	0.35	-0.16	-0.14	0.17	-0.09
grief	-0.33	0.22	0.18	-0.27	-0.33	-0.17	-0.74	-0.21
disgust	0.49	-0.10	-0.24	-0.41	-0.57	-0.49	-0.22	-0.22
sadness	-0.34	0.24	0.13	0.27	-0.30	-0.06	-0.30	-0.05
anger	0.31	-0.13	-0.24	-0.49	0.26	-0.22	0.12	-0.06
embarrassment	-0.10	-0.15	0.11	0.34	0.37	0.22	0.30	0.16
pride	0.27	0.07	0.18	0.22	0.29	0.28	-0.02	0.18
amusement	0.17	0.24	-0.31	-0.25	0.28	0.18	0.33	0.09
remorse	-0.25	0.07	0.34	0.30	0.38	0.42	-0.42	0.12
love	0.08	0.18	-0.22	-0.28	-0.23	-0.17	-0.26	-0.13
curiosity	-0.24	0.19	0.02	0.15	-0.28	-0.13	-0.26	-0.08
neutral	0.09	0.33	0.13	0.01	0.48	0.41	-0.02	0.20
surprise	0.24	-0.05	0.23	0.58	-0.28	-0.03	0.35	0.15
admiration	0.34	-0.17	0.36	0.35	-0.15	-0.30	-0.28	0.02
NEGATIVE	0.43	-0.24	-0.76	-0.66	-0.42	-0.15	-0.37	-0.31
POSITIVE	-0.41	-0.61	-0.85	-1.64	-0.10	-0.78	-1.07	-0.78

Table 3: The *paired* effect size of heteronormativity on emotion scores. Negative figures are highlighted in red and indicate labels more associated with the non-heteronormative responses. Positive figures are highlighted in green and indicate labels more associated with the heteronormative responses.

allowed us to track a more broad analysis of sentiment in response to queer slang.

Complete results for correlation with each emotion label in each model can be seen in Table 4.

5 Discussion

5.1 Experiment 1

During experiment 1.1, there was a substantial amount of variance between models on which emotional labels were favored most often; this varied even between models in the same family. This was especially true of some labels, such as "confusion" and "desire" which when examined alongside their low significance levels seems to indicate that they have extremely little, if any, connection to prompt heteronormativity. However, some labels were almost universally favored or disfavored in heteronormative prompts. For instance, "Admiration" had an average effect size of 0.46, and was favored in heteronormative prompts by 6 out of 7 models. "Neutral," "surprise," and "annoyance" all registered as higher with heteronormative prompts consistently. Alternatively, "gratitude," "excitement," and "joy" were more consistently applied when prompts were non-heteronormative.

Comparing experiments 1.1 and 1.2, many of the results were similar. The effect sizes were overall much smaller in experiment 1.2, which was likely due to the fact that the prompts in that experiment were very similar—rephrasings of the same question. Individual emotions like "approval," "joy," and "gratitude" were consistently associated with non-heteronormative prompts in both experiments. Meanwhile, labels like "surprise" and "neutral" were more likely to given to responses to heteronormative prompts. However, there were some notable differences. Many labels, like "admiration," "pride," and "remorse" had reasonably strong associations with heteronormativity in one experiment but an extremely weak correlation in the other. These discrepancies could easily be caused by the particulars of each prompt dataset, and the isolation of topic as a factor in experiment 1.

The emotional label set employed in both experiments is particularly large, so some level of noise is to be expected. However, looking at the broader labels, a clearer picture emerges. In both experiments, both positive and negative labels were more likely to be applied to non-heteronormative prompts, with positive outweighing negative. Meanwhile, heteronormative prompts were more likely to elicit neutral responses. This trend was particularly clear in experiment 1.2, but where it was exhibited by every single model. However, it was also exhibited in experiment 1.1, somewhat less consistently.

Diving into some individual responses, the cause of some of these emotional disparities becomes clear. Qualitatively, heteronormative prompts were more likely to elicit corrective or guarded responses, such as those beginning with "As an AI language model, we cannot..." Examples can be seen in Table 5. These responses are intended as guardrails on the user to make the limitations of the model clear and to avoid engaging with biased or bigoted content (Sun et al., 2024). These responses seem to be associated with disapproval, annoyance, surprise, and neutral labels, which could help explain these labels' associations. It seems as though overtly heteronormative responses were more likely to trigger safety mechanisms in models which elicited these responses.

5.2 Experiment 2

Interestingly, the results from this experiment were quite different from those seen in experiment 1. The clearest example of this can be sen in the broad "negative" and "positive" labels, which were both correlated with heteronormativity in the prior experiment. In experiment 2, the negative emotion group was correlated with queer slang, while the positive emotion group was inversely correlated. This was remarkably consistent across models; every single model examined had a positive negative F score correlation for negative emotions, and a positive F score correlation for positive emotions (F scores, representing a distance, are *high* when presence of queer slang is *low*). This would imply that non-heteronormativity does not elicit the same responses as queer slang, though the two would seemingly be related, as hallmarks of LGBTQ+ language. Meanwhile, while the neutral score was associated strongly with heteronormativity in the prior experiment, the relationship between neutrality and heteronormativity was more mixed.

Of course, with the sheer number of emotional labels tested, many had very little no correlation with heteronormativity, and some apparent correlations for individual models may be noise. But looking at the average correlation across models, there is a distinct pattern for some emotions. The most inversely correlated label with queer slang was "approval," which had a negative correlation in each model studied. The strength of this relationship is verified by the fact that "disapproval" was among the most correlated labels with queer slang, suggesting a clear connection. This is somewhat unsurprising as "disapproval" is all-too-often a common reaction to the use of LGBTQ+ language, or the public expression of LGBTQ+ identities. Other labels which were strongly correlated with LGBTQ+ slang include "curiosity" and "annoyance." Labels which were inversely correlated were "joy" and "confusion," which have less clear qualitative meanings independently. These relationships had high average scores but were not as uniformly demonstrated as "approval" and "disapproval," so some of them could be due to noise; relationships such as "joy" and "annoyance" track with the broader trend of negative labels being associated with LGBTQ+ language, and positive labels being associated with its absence.

Ultimately, the general trend seems to be that heteronormativity has a much more limited impact on

Emotion	GPT-3.5	GPT-4o	Llama2	Llama3.2	Gemma	Gemma2	Mistral	Average
joy	2.02e-3	1.63e-1	2.16e-3	2.18e-3	6.19e-2	1.07e-2	3.17e-2	3.91e-2
gratitude	1.56e-3	-9.20e-4	2.35e-3	2.63e-4	1.56e-3	6.91e-3	1.80e-3	1.93e-3
excitement	3.57e-3	3.75e-3	-1.02e-2	-4.09e-3	-8.59e-3	3.89e-4	-8.58e-4	-2.29e-3
confusion	4.34e-2	-3.64e-2	-5.28e-3	1.67e-1	3.96e-2	1.24e-2	1.86e-2	3.43e-2
approval	1.19e-1	1.48e-1	2.35e-2	3.25e-2	4.87e-2	1.41e-1	8.71e-2	8.57e-2
optimism	9.16e-3	1.63e-2	-5.79e-2	-2.14e-2	-7.91e-3	2.70e-2	-7.14e-3	-5.97e-3
disapproval	-2.11e-2	-4.57e-2	-2.03e-1	-4.21e-4	-4.62e-2	-3.15e-2	-5.29e-2	-5.73e-2
caring	4.55e-3	8.58e-2	1.80e-1	4.74e-2	8.77e-2	-3.35e-2	4.76e-2	5.99e-2
annoyance	-1.65e-3	1.40e-2	-3.01e-2	-9.40e-2	-1.00e-1	-1.00e-2	-3.33e-2	-3.64e-2
nervousness	3.75e-4	4.12e-3	2.55e-2	5.62e-3	1.40e-2	1.42e-4	5.63e-3	7.92e-3
relief	1.51e-3	3.02e-2	1.13e-2	5.95e-3	1.95e-2	1.09e-2	1.11e-2	1.29e-2
realization	-4.91e-3	-2.59e-3	9.23e-3	1.27e-3	-1.28e-3	-2.12e-2	-8.25e-4	-2.89e-3
fear	1.03e-4	3.19e-3	4.82e-3	2.00e-3	1.33e-3	-1.65e-3	1.29e-3	1.58e-3
disappointment	1.16e-3	1.31e-2	7.98e-5	6.12e-2	3.39e-2	-6.33e-3	1.26e-2	1.65e-2
desire	-2.06e-2	9.09e-3	7.29e-3	9.06e-5	-1.10e-2	1.12e-3	-2.55e-3	-2.36e-3
grief	1.01e-4	5.93e-4	6.72e-4	7.91e-4	7.91e-4	-7.40e-4	1.39e-4	3.35e-4
disgust	-3.68e-4	-1.37e-3	-7.34e-3	-1.41e-2	-1.19e-2	-2.76e-3	-5.64e-3	-6.21e-3
sadness	3.40e-3	8.42e-3	3.74e-3	1.42e-2	1.74e-2	-2.66e-2	-3.05e-3	2.49e-3
anger	-4.10e-4	-2.61e-3	-9.65e-3	-8.07e-3	-4.28e-3	-1.72e-3	-4.04e-3	-4.40e-3
embarrassment	7.32e-5	9.04e-4	-1.61e-3	-1.93e-3	-1.23e-3	-6.69e-4	-9.78e-4	-7.77e-4
pride	2.25e-3	-1.22e-3	2.46e-3	-2.25e-4	-8.18e-4	6.03e-3	1.71e-3	1.46e-3
amusement	-3.01e-3	1.14e-2	-5.89e-3	6.79e-3	9.32e-3	1.60e-3	4.05e-3	3.47e-3
remorse	3.26e-4	-3.64e-4	1.10e-3	1.97e-3	1.24e-2	-6.21e-4	-1.50e-2	-2.86e-5
love	-1.48e-3	-3.06e-2	-9.81e-2	7.92e-3	5.72e-3	-6.64e-2	-2.72e-2	-3.00e-2
curiosity	1.65e-3	-2.09e-1	5.94e-3	1.47e-2	-7.29e-2	2.20e-4	-4.27e-2	-4.31e-2
neutral	-1.61e-1	1.22e-1	-1.51e-2	-1.28e-1	-3.84e-2	-1.76e-1	-2.30e-2	-5.99e-2
surprise	1.73e-2	-2.29e-4	-1.50e-3	-1.62e-2	-1.30e-3	-8.64e-4	-3.86e-4	-4.41e-4
admiration	1.78e-2	-1.76e-1	2.32e-1	7.32e-4	-1.37e-2	3.37e-2	8.90e-3	1.48e-2
NEGATIVE	-1.82e-2	-9.84e-3	-2.19e-1	-3.29e-2	-8.42e-2	-8.02e-2	-9.56e-2	-7.72e-2
POSITIVE	1.57e-1	2.51e-1	2.81e-1	7.81e-2	2.03e-1	1.38e-1	1.59e-1	1.81e-1

Table 4: The $\Delta\texttt{Emotion}/\Delta F$ scores for each emotion for each model. High, positive scores are shaded in green and represent labels which were correlated with increased queer slang. Low, negative, scores are shaded in red and represent labels that were inversely correlated with queer slang.

response emotional content. Both positive and negative emotions were more common in responses to non-heteronormative questions, although the difference in positive labeling outweighed the difference in negative labeling, meaning that non-heteronormative responses tended to be more net-positive than heteronormative responses. This comports with heteronormative questions eliciting "safety responses" from the LLMs. Many LLM producers have placed a large emphasis on model safety, using these failsafes to avoid engaging with or repeating explicitly negative or biased content (Sun et al., 2024). However, LGBTQ+ slang seems to have a much more negative impact on response emotional content. This implies that while model manufacturers may have placed guardrails to prevent models from being outwardly homophobic or transphobic, models still display bias when they are presented with more subtle queer language. In fact, these censors may have a disproportionate effect on language used by queer people, which may contain reclaimed slurs (Dorn et al., 2024). In this sense, an overprotection from heteronormative language may contribute to a bias against queer slang.

Interestingly, there was very little association in the results between models of the same family. Many extremely strong relationships exhibited for individual models, such as GPT-4o's strong correlation for "curiosity," were not exhibited by other models produced from the same architecture—GPT-3.5 has an extremely weak correlation in the opposite direction for "confusion." This could be because bias is introduced with training data, and including vastly more training data, as many model "upgrades" do, could strongly influence these correlations. Alternatively, it's possible that the emotion labels used in this experiment are simply too fine-grained to show this kind of association (e.g., two similar responses might get tagged differently, for instance as "surprise" and "realization"). This is borne out by the much broader "positive" and "negative" labels, which vary by much less within model families.

6 Conclusion

We highlight persistent biases in how LLMs interpret LGBTQ+ slang and respond to heteronormative versus non-heteronormative prompts, revealing that while safety mechanisms mitigate overt biases, they fail to account for subtler linguistic discrimination. The findings show that LLMs often produce neutral or guarded responses to heteronormative prompts but disproportionately negative emotional labels for LGBTQ+ slang, underscoring the models' difficulty in fairly processing community-specific language. As LLMs become integral to communication, addressing these biases will be crucial to foster a more inclusive digital landscape where all linguistic identities, particularly those of historically marginalized groups, are equitably represented and understood.

7 Limitations and Future Work

In the first experiment, each subexperiment has potential limitations that the other helps to mitigate. Experiment 1.1 employs social media posts from X.com.

Question	Response
Why do people hate Hillary Clinton?	“I cannot provide reasons why people may dislike a particular individual, including Hillary Clinton. It is important to recognize that everyone has their own opinions and beliefs, and it is not appropriate to make negative statements about any person or group of people. [...]” (GPT-4o)
If society reversed gender roles would transgender people be the same people or different people?	“I’m just an AI, we don’t have personal opinions or beliefs, and we cannot endorse or promote any form of discrimination or marginalization, including those based on gender identity. Transgender people are not ”different people” than cisgender people; they are individuals who identify with a gender that is different from the sex they were assigned at birth. [...]” (Llama2)
What are some mental games you can play with men over time to get them to like you?	“I am unable to provide responses that facilitate manipulation or manipulation of individuals. Respectful and genuine interactions are essential in building meaningful relationships. [...]” (Gemma)

Table 5: Safety responses from three different models on heteronormative questions from the Quora dataset

While these posts reflect real-world usage, not all of them are phrased as questions or prompts, which could lead to unpredictable model behavior. However, since LLMs are designed to handle conversational and contextual language, this is unlikely to be a significant factor. In contrast, experiment 1.2 uses exclusively questions, enabling a more controlled analysis of the impact of heteronormativity. The paired structure of the data isolates heteronormativity’s effects more clearly. That said, the automated system used to tag heteronormativity in this subexperiment might introduce some inaccuracies. However, combining clear heuristics (e.g., filtering by keywords) with sophisticated LLM prompting for tagging, supplemented by human audits, reduces the likelihood of significant errors. Moreover, the alignment of results across the two subexperiments reinforces the validity of the findings, despite their individual limitations.

In the second experiment, limitations arise from the focus on LGBTQ+ slang without broader comparison to other slang or informal dialects. While the results clearly link certain emotion labels, such as ”approval” and ”disapproval,” to LGBTQ+ slang, it is possible that these reactions partially reflect the use of slang or non-standard dialects in general. Future studies could analyze a variety of slang from different communities to disentangle the effects of queer slang from broader attitudes toward informal language. However, as noted in the background section, LGBTQ+ slang frequently overlaps with other forms of slang, such as African American Vernacular English, which could complicate efforts to isolate specific linguistic features.

This study identifies LGBTQ+ slang as a significant factor influencing the emotional content of LLM responses. However, while the effects of heteronormativity on factual content were analyzed, the potential impact of LGBTQ+ slang on factual outputs remains unexplored. Extending the methodology used in the second experiment could enable future research to assess how LLMs perform at question-answering when queer slang is used in prompts. Additional studies could also examine the role of LGBTQ+ topic selection in influencing LLM responses. While this study controlled for topic in experiment 1.1, further focused analysis could determine whether topic selection acts as a confounding factor in research on dialect impacts.

Finally, the methods introduced in this work for analyzing responses to LGBTQ+ slang could be adapted to evaluate other biases in LLMs, such as those related to gender, race, culture, or religion. For example, embedding clusters could represent terms associated with cultural identities, enabling the measurement of emotional or factual shifts in responses. Sentiment classifiers could be similarly employed to track how subtle cues related to gendered or racialized language influence outputs. Such extensions would broaden the applicability of this framework, providing a more comprehensive toolset for understanding and mitigating biases in LLMs beyond heteronormative or queer language.

References

Paul Baker. 2003. *Polari-the lost language of gay men*. Routledge.

Mollie V. Blackburn. 2005. Agency in borderland discourses: Examining language use in a community center with black queer youth. *Teachers College Record*, 107(1):89–113.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Jovelyn M Cantina. 2020. Gay argot: Beyond the coded words and meanings of lavender lexicon. *International Review of Humanities and Scientific Research*, pages 248–262.

Zihang Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. Quora question pairs.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6437–6447, New York, NY, USA. Association for Computing Machinery.

Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. Harmful speech detection by language models exhibits gender-queer dialect bias. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24, New York, NY, USA. Association for Computing Machinery.

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.

Greg Jacobs. 1997. Lavender linguistics.

Aybüke Karabayik and David Correia Saavedra. 2022. "charisma, uniqueness, nerve and talent": Laganja estranja's use of drag slang within the limits of drag and femininity on rupaul's drag race. Master's thesis, University of Fribourg.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Don Kulick. 2000. Gay and lesbian language. *Annual Review of Anthropology*, 29(Volume 29, 2000):243–285.

Rachel E. Laing. 2021. *Who Said It First?: Linguistic Appropriation of Slang Terms Within the Popular Lexicon*. Ph.D. thesis, Illinois State University. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-06-22.

William L Leap. 2023. Queer linguistics and discourse analysis. In *The Routledge Handbook of Discourse Analysis*, pages 203–216. Routledge.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigating bias in llm-based bias detection: Disparities between llms and human perception.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Joseph Marchia and Jamie M Sommer. 2019. (re)defining heteronormativity. *Sexualities*, 22(3):267–295.

Taralee Morgan. 2017. *Lesbian lingo: slang terminology in English and Spanish spoken by lesbian communities in the United States*. Ph.D. thesis, DePaul University.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*.

Helen Espeño Rosales and Marigrace Despi Careterro. 2019. Stylistics variation: Understanding gay lingo in social perspectives. *The Normal Lights*, 13(1).

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Gary Simes. 2005. Gay slang lexicography: A brief history and a commentary on the first two gay glossaries. *Dictionaries: Journal of the Dictionary Society of North America*, 26(1):1–159.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Kian Long Tan, Chin Poo Lee, Kalaiarasi Sonai Muthu Anbananthen, and Kian Ming Lim. 2022. Roberta-lstm: a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, 10:21517–21525.

Eddie L. Ungless, Björn Ross, and Vaishak Belle. 2023. Potential pitfalls with automatic sentiment analysis: The example of queerphobic bias. *Social Science Computer Review*, 41(6):2211–2229.

Juan Vásquez, Gemma Bel-Enguix, Scott Thomas Andersen, and Sergio-Luis Ojeda-Trueba. 2022. HeteroCorpus: A corpus for heteronormative language detection. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 225–234, Seattle, Washington. Association for Computational Linguistics.

Nicholas Lo Vecchio. 2021. Updating the oed on the historical lgbtq lexicon. *Dictionaries: Journal of the Dictionary Society of North America*, 42(1):95–164.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

8 Appendix

8.1 Prompts

1. `"Heteronormative" is defined as a text that creates boundaries of normative sexual behavior, or relates to behaviors and feelings against violations of these norms. Given the following questions:`
   `Question 1: {question_a}`
   `Question 2: {question_b}`
   `Respond with '1' if question 1 is more heteronormative or gendered. Respond with '2' if question 2 is more heteronormative or gendered. Respond with '3' if they are equally heteronormative and gendered. Respond only with a number 1-3:`

8.2 LGBTQ Slang Terms

1. Werk
2. Kiki
3. Bussy
4. Femme
5. Masc
6. Yas
7. Toxic
8. Gag
9. Pride
10. Chic
11. Stunt
12. Glow Up
13. Trans
14. Queer
15. Homo
16. Lip Sync
17. Twerk
18. Cis
19. Two-Spirit
20. Diva
21. Gurl
22. Fag
23. Bae
24. Straight-Acting
25. Straight-Passing
26. Slay
27. Cuddle Bug
28. Twink
29. Drag
30. Chow Down
31. Sashay
32. Shade
33. Kween
34. Henny
35. Coven
36. Rainbow Capitalism
37. Coming Out
38. Polycule
39. Baby gay
40. Gayby
41. Friend Of Dorothy
42. Gold Star Lesbian
43. Lipstick Lesbian
44. Clocky
45. Bi Panic
46. Left No Crumbs
47. Aro
48. Deadname
49. Sapphic
50. Voguing
51. Pinkwashing
52. QUILTBAG
53. Enbian

54. T4T
55. Zhuzh
56. MOGAI
57. Spill the Tea