Leveraging Large Language Models in Detecting Anti-LGBTQIA+ User-generated Texts

Quoc-Toan Nguyen[1], **Josh Nguyen**[2], **Van-Tuan Pham**[3], **William John Teahan**[4]

[1] University of Technology Sydney, [2] The University of Melbourne,
[3] RMIT University, [4] Bangor University

🖂Corresponding Author: w.j.teahan@bangor.ac.uk

Abstract

Anti-LGBTQIA+ texts in user-generated content pose significant risks to online safety and inclusivity. This study investigates the capabilities and limitations of five widely adopted Large Language Models (LLMs)—DeepSeek-V3, GPT-4o, GPT-4o-mini, GPT-o1-mini, and Llama3.3-70B—in detecting such harmful content. Our findings reveal that while LLMs demonstrate potential in identifying offensive language, their effectiveness varies across models and metrics, with notable shortcomings in calibration. Furthermore, linguistic analysis exposes deeply embedded patterns of discrimination, reinforcing the urgency for improved detection mechanisms for this marginalised population. In summary, this study demonstrates the significant potential of LLMs for practical application in detecting anti-LGBTQIA+ user-generated texts and provides valuable insights from text analysis that can inform topic modelling. These findings contribute to developing safer digital platforms and enhancing protection for LGBTQIA+ individuals.

⚠Warning: Given the research's objectives, this paper includes profanity, vulgarity, and other harmful language. These may be disturbing for queer or LGBTQIA+ individuals and other readers.

1 Introduction

The dramatic growth of user-generated content (Gorwa et al., 2020) underscores the urgent need to prevent the spread of intentionally and unintentionally harmful material across Online Social Networks (OSNs) or other digital platforms. Initially, user-generated text moderation relied on manual, rule-based methods, but with advancements in Artificial Intelligence (AI), OSNs and digital platforms have increasingly applied advanced technologies to uphold platform integrity. These developments are essential to protect both users and online communities from harmful content (Franco et al., 2024).

Abusive language or cyberbullying are among the most vital problems, and continue to pose significant challenges worldwide, affecting a vast number of individuals (Hong et al., 2025). If left unaddressed, such harmful interactions can greatly heighten the risk of suicidal thoughts and behaviours (Gini and Espelage, 2014). Although the relationship between bullying and suicidality—including suicidal ideation and attempts—is complex, research strongly indicates that victimization plays a major role in increasing this risk, often leading to severe psychological consequences for those affected (Holt et al., 2015). A promising approach to mitigating this problem is the development of AI-based moderation systems (Cedric et al., 2022; Todor et al., 2023; Calabrese et al., 2024), which can efficiently detect abusive language on a large scale. Especially, utilising **Large Language Models** (**LLMs**) has notably advanced this task (Neele et al., 2024; Sarah et al., 2024; Prince et al., 2024; Franco et al., 2024; Wei et al., 2024; Hyundong et al., 2024).

However, the previous studies typically adopt a universal framework neglecting the evaluation and specific development, leading to high potential risks for queer individuals (Jordan et al., 2024) or LGBTQIA+ community (Are et al., 2024) [1] despite growing evidence that they experience cyberbullying at significantly higher rates and at significantly higher rates than their heterosexual peers (Oliver et al., 2021; Abreu and Kenny, 2018). Cyberbullying among LGBTQIA+ individuals has been linked to a wide range of harmful consequences (Abreu and Kenny, 2018), including severe psychological and emotional distress such as depression, low self-esteem, and an increased risk of suicidal thoughts and attempts. Furthermore, it can also contribute

[1] LGBTQIA+ stands for lesbian, gay, bisexual, transgender, queer or questioning, intersex, and asexual. The "+" symbol includes other identities that may not be explicitly listed in the acronym.

to behavioural issues, such as heightened physical aggression, body image concerns, and social isolation (Abreu and Kenny, 2018). Therefore, when leveraging LLMs for user-generated text moderation on OSNs or other digital platforms (websites, mobile apps,...), it is crucial to assess their effectiveness in identifying harmful or anti-LGBTQIA+ user-generated text (Schey and Shelton, 2023) before deployment. Without proper evaluation, LLMs may fail to recognize subtle forms of discrimination, reinforce biases, or even inadvertently allow harmful user-generated texts to persist, ultimately exacerbating the challenges faced by people in the LGBTQIA+ community.

Hence, in this paper, we leverage five LLMs including DeepSeek-V3 (Liu et al., 2024), GPT-4o (Hurst et al., 2024), GPT-4o mini (Hurst et al., 2024), GPT-o1-mini (Aaron et al., 2024), and Llama3.3-70b (Jonas et al., 2025) which are among the most widely-used and up-to-date methods in the literature, to answer the following Research Questions (RQs) using user-generated texts comments data from YouTube, Reddit, and X with anti-LGBTQIA+ user-generated content (Pratik et al., 2022):

- **RQ1**: What are the predominant linguistic patterns and strongest associations in anti-LGBTQIA+ user-generated texts?

- **RQ2**: How can we leverage LLMs to detect anti-LGBTQIA+ user-generated texts?

- **RQ3**: How effectively do LLMs detect anti-LGBTQIA+ user-generated texts, and how do their predictive performance and calibration differ in this task?

2 Related Work

Recent studies highlight the growing role of LLMs in automated content moderation. Sarah et al. (2024) proved LLMs' contextual understanding aids hate speech detection. Neele et al. (2024) proved the potential of user-driven moderation but pointed out scalability challenges. Wei et al. (2024) demonstrated that LLM pipelines reduce computational costs while maintaining high accuracy. Cedric et al. (2022) emphasized detecting minority arguments for better understanding in debates. Franco et al. (2024) highlighted LLMs' support in moderation dynamics, though reasoning limitations remain. Kou and Gui (2020) stressed the importance of community-aligned explanations in AI-led moderation.

Nevertheless, existing research predominantly addresses the general population, with limited evaluation of these methods specifically for the LGBTQIA+ community. This gap raises concerns regarding potential biases and shortcomings that may disproportionately affect this marginalised group. However, studies addressing this issue remain limited in the literature. While LLMs have shown advancements over conventional AI models, current methodologies for anti-LGBTQIA+ user-generated content still primarily rely on conventional AI approaches (Vivek et al., 2024; Arora et al., 2024).

3 Methods

3.1 Material

This study utilises a part of a dataset by Pratik et al. (2022) comprising social media comments collected from various users on YouTube, Reddit, and X (formerly known as Twitter). They were labelled by 11143 annotators recruited via Amazon Mechanical Turk. 4299 samples are manually selected and labelled from the original dataset, specifically focusing on content relevant to anti-LGBTQIA+ research following these references' approaches of how to curate the data (J et al., 2024). The data is published by Patel (2025).

We categorise the dataset into two groups: non-anti-LGBTQIA+ and anti-LGBTQIA+. In its raw form, the non-anti-LGBTQIA+ category contains 109764 words, 9634 unique words, while the anti-LGBTQIA+ category comprises 21157 words, with 4023 unique words. After applying common stop words removal, as per (Nothman et al., 2018), the word count for non-anti-LGBTQIA+ is reduced to 50243 words with 9471 unique words (3043 samples), whereas the anti-LGBTQIA+ category retains 11072 words with 3867 unique words (1256 samples). This preprocessing step may refine the dataset for further text analysis. Figure 1 presents examples from both classes.

3.2 User-generated Anti-LGBTQIA+ Text Detection using Large Language Models

Figure 2 illustrates our workflow, demonstrating how LLMs are leveraged for anti-LGBTQIA+ text detection using an example user, x. The process begins when user x sends a text input, denoted as x_{text}. This input is then incorporated into a

Non-anti-LGBTQIA+

- The trans women reading this tweet right now is beautiful
- Sounds like the kinda wholesome life I'd die for never met any black Lesbians before so i guess you're my first ever.
- I met my first gay person, one of my brother's friends, when I was 14. I remained deeply invested in my Methodist faith until I was about 17. I knew the rule—'Man shall not lie with man'—but Jesus also taught to love everyone. So, I wrestled with the question.
- I never switched to bi because of pressure. But I just wanted to leave some happiness in your comments. Don't feel pressured to change just to make things easier for others. Identify as whatever makes you happy and comfy. You're awesome, so be your awesome self! That way you'll be happy and spread your happiness to all of your loved ones.

Anti-LGBTQIA+

- Imagine unironically watching that braindead tranny.
- Quit fucking being an ableist you faggot. Anger issues can't just be dealt with. Kill yourself you pot smoking brain dead one-celled amoeba. Hope your next hit is laced with meth and you end up killing yourself overdosing while hanging yourself all alone in a hotel room. Fucking poser ass retard.
- Women football is trash .. Rapinoe is a filthy lesbian .. Tel her to repent
- And I don't care if you're gay, as long as you aren't a FILTHY CUMSUCKING FAGGOT about it. There is a difference. Don't make being gay your ENTIRE EXISTENCE, and constantly talk about your sex life as if that's the sum of who you are as a person. Not everyone who doesn't agree with homosexual is am evil bigot who wants to herd you all into death camps and electrocute you into being straight.

Figure 1: Examples of user-generated texts from the dataset within the two classes.

developed prompt as x_{input}.

In this prompt, we utilise a “zero-shot" approach (Li et al., 2024; Pengyue et al., 2024; Chi et al., 2024). It is a technique in Natural Language Processing (NLP) where a model performs a task without being provided with specific examples related to that task (Tom et al., 2020; Ross et al., 2023; Hu et al., 2024). Rather than learning from explicit demonstrations, the model relies on a direct task description within the prompt, utilising its pre-trained knowledge and reasoning abilities to generate an appropriate response. This approach enables models to adapt to various tasks without requiring additional fine-tuning.

In this prompt, various LLMs—DeepSeek-V3 (Liu et al., 2024), GPT-4o (Hurst et al., 2024), GPT-4o mini (Hurst et al., 2024), GPT-o1-mini (Aaron et al., 2024), and Llama3.3-70b (Jonas et al., 2025)—is leveraged to function as moderator(s) to analyse the given text. Each model represented as F, processes x_{input} to provide an output, denoted as y_{output}, determining whether the text is classified as anti-LGBTQIA+. Additionally, the model provides a score, c, indicating the confidence of its prediction following this black-box approach, asking the confidence score directly from the prompts (Youliang et al., 2024).

For the LLMs' evaluation, multiple samples from various users—denoted as $(x_{\text{text}}^{(1)}, x_{\text{text}}^{(2)}, \ldots, x_{\text{text}}^{(N)})$—are collected from the material described in Section 3.1. Each sample is sequentially processed by the LLMs, including DeepSeek-V3 (Liu et al., 2024), GPT-4o (Hurst et al., 2024), GPT-4o mini (Hurst et al., 2024), GPT-o1-mini (Aaron et al., 2024), and Llama3.3-70b (Jonas et al., 2025). Each sample has an individual classification result y and confidence score c. The final evaluation aggregates the results across all processed samples, ensuring a comprehensive assessment of model performance. The entire workflow for anti-LGBTQIA+ text detection using LLMs can be summarised mathematically as:

$$Y = F(X) = \{(y_i, c_i) \mid y_i, c_i = F_j(x_i), \forall x_i \in X, \forall F_j \in \mathcal{F}\}$$

where:

- $X = \{x_{\text{text}}^{(1)}, x_{\text{text}}^{(2)}, \ldots, x_{\text{text}}^{(N)}\}$ represents the set of user text inputs.
- F_j is an LLM from the set of models:

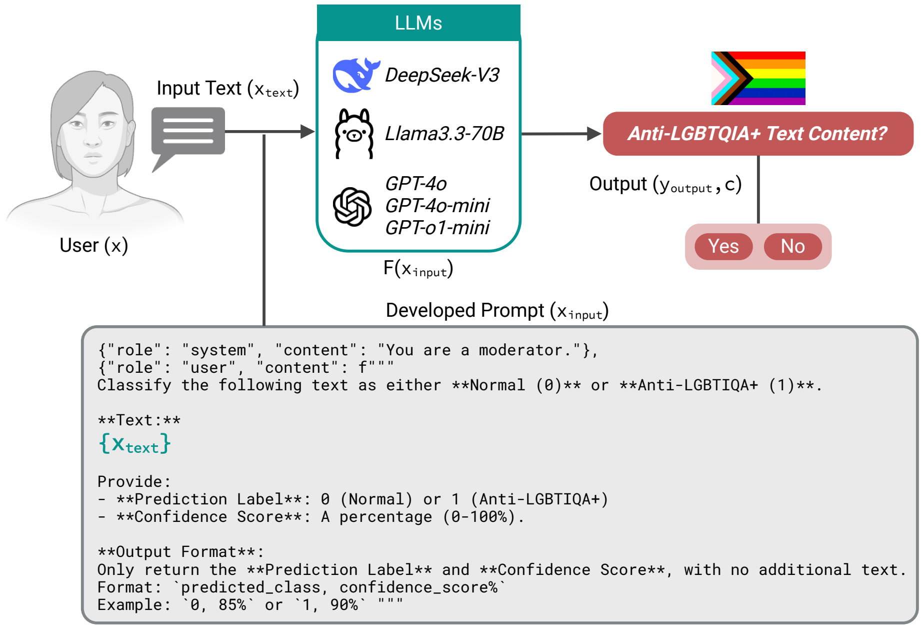
  $$\mathcal{F} = \{\text{DeepSeek-V3}, \text{Llama3.3-70b}, \text{GPT-4o}, \text{GPT-4o mini}, \text{GPT-o1-mini}\}$$

- Each model F_j takes an input x_i (transformed into $x_{\text{input}}^{(i)}$ through prompting) and outputs:
  - $y_i \in \{0, 1\}$, where 1 indicates the text is classified as anti-LGBTQIA+ and 0 otherwise.
  - $c_i \in [0, 1]$, the confidence score of the classification.

4 Experiments

The experiments of LLMs in this research are completed via model APIs provided by Open AI (OpenAI, 2025) (GPT-4o (Hurst et al., 2024), GPT-4o mini (Hurst et al., 2024), and GPT-o1-mini (Aaron et al., 2024)), and Meta Llama (Meta, 2025) (Llama3.3-70b (Jonas et al., 2025), DeepSeek-V3 (Liu et al., 2024)). The default hyperparameters are set, including `temperature=1.0`, `Top_p=1.0`, and `presence_penalty=0.0`.

In text analysis, a word cloud (Jin, 2017) is used to visualize the top 30 most frequent words in anti-LGBTQIA+ user-generated texts. Additionally, the strongest associations between commonly occurring offensive terms are analysed and visualised by



```
{"role": "system", "content": "You are a moderator."},
{"role": "user", "content": f"""
Classify the following text as either **Normal (0)** or **Anti-LGBTIQA+ (1)**.

**Text:**
{x_text}

Provide:
- **Prediction Label**: 0 (Normal) or 1 (Anti-LGBTIQA+)
- **Confidence Score**: A percentage (0-100%).

**Output Format**:
Only return the **Prediction Label** and **Confidence Score**, with no additional text.
Format: `predicted_class, confidence_score%`
Example: `0, 85%` or `1, 90%` """
```

Figure 2: Workflow of the method used for leveraging Large Language Models (LLMs) for user-generated anti-LGBTQIA+ text detection.

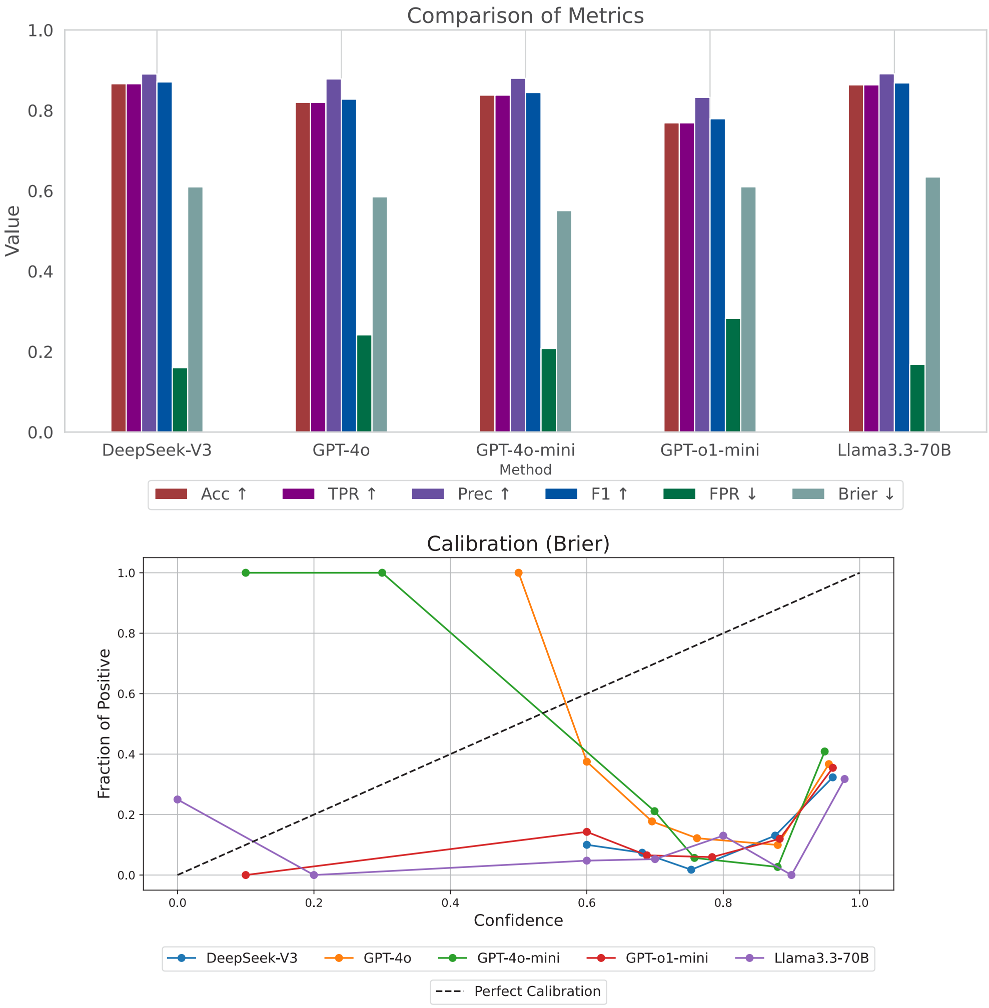


Figure 3: Performance of Large Language Models (LLMs) with metrics. Accuracy ($Acc \uparrow$), True Positive Rate ($TPR \uparrow$), Precision ($Prec \uparrow$), F1-score ($F1 \uparrow$), False Positive Rate ($FPR \downarrow$) and Brier score ($Brier \downarrow$).

a bigram network graph, which visualizes the top 30 most frequent bigrams, using a library named Networkx (Hagberg and Conway, 2020).

Regarding evaluation metrics, five key performance indicators are utilised, each playing a crucial role in AI applications (Hicks et al., 2022). These metrics include **Accuracy** ($Acc \uparrow$), **True Positive Rate** ($TPR \uparrow$), **False Positive Rate** ($FPR \downarrow$), **Precision** ($Prec \uparrow$), and **F1-score** ($F1 \uparrow$). Additionally, the Brier score ($Brier \downarrow$) (Rufibach, 2010) is incorporated as a metric of probabilistic calibration (Youliang et al., 2024). The values for TPR, FPR, Prec, and F1 are computed using macro-averaging. These metrics range from 0 to 1, where **higher values correspond to better performance** for all metrics, except for FPR and $Brier$, where a **better model has lower values**.

5 Results

Word Frequency			Association Frequency	
Word	**Percentage**	**Count**	**Association**	**Count**
faggot	3.775	418	(’fucking’, ’faggot’)	33
fuck	1.933	214	(’suck’, ’dick’)	24
fucking	1.852	205	(’fuck’, ’faggot’)	23
gay	1.581	175	(’shut’, ’fuck’)	21
ass	1.228	136	(’faggot’, ’ass’)	18
faggots	1.030	114	(’piece’, ’shit’)	18
shit	1.021	113	(’gon’, ’na’)	16
fag	0.939	104	(’wan’, ’na’)	13
bitch	0.912	101	(’fuck’, ’gay’)	12
dick	0.894	99	(’mentally’, ’ill’)	11
shut	0.560	62	(’gay’, ’ass’)	10
suck	0.551	61	(’ass’, ’bitch’)	10
fags	0.470	52	(’eat’, ’shit’)	10
stupid	0.434	48	(’shut’, ’faggot’)	9
kill	0.415	46	(’ass’, ’faggot’)	9
hate	0.334	37	(’gay’, ’shit’)	9
retarded	0.316	35	(’fucking’, ’gay’)	9
tranny	0.316	35	(’burn’, ’hell’)	8
pussy	0.307	34	(’retarded’, ’faggot’)	7
queer	0.307	34	(’fuck’, ’fag’)	7
dumb	0.289	32	(’fuck’, ’fucking’)	7
die	0.289	32	(’dick’, ’die’)	7
homosexual	0.271	30	(’child’, ’molester’)	7
retard	0.271	30	(’pussy’, ’ass’)	7
god	0.262	29	(’shit’, ’faggot’)	7
ill	0.244	27	(’fucking’, ’bitch’)	6
cunt	0.244	27	(’bunch’, ’faggots’)	6
hell	0.235	26	(’baby’, ’raping’)	6
disgusting	0.235	26	(’fucking’, ’faggots’)	6
cock	0.235	26	(’stupid’, ’fucking’)	6

Table 1: Top 30 most frequent words and strongest bigram associations in anti-LGBTQIA+ user-generated texts.

5.1 Analysis of Words in Anti-LGBTQIA+ User-generated Texts

To begin with, Table 1 and Figure 4 present the most frequently occurring words in anti-LGBTQIA+ user-generated texts (stop words removed), highlighting derogatory language and hate speech. The most common term, “faggot,” accounts for 3.775% of occurrences (418 times), followed by profanity such as “fuck” (1.933%, 214 times) and “fucking” (1.852%, 205 times). Several slurs targeting LGBTQIA+ individuals, including “gay,” “faggots,” “fag,” “tranny,” and “queer,” appear with notable frequency. Additionally, the presence of words associated with aggression (e.g., “kill,” “hate,” “die”) and derogatory terms like “retarded” and “stupid” further underscores the negative sentiment in these texts.

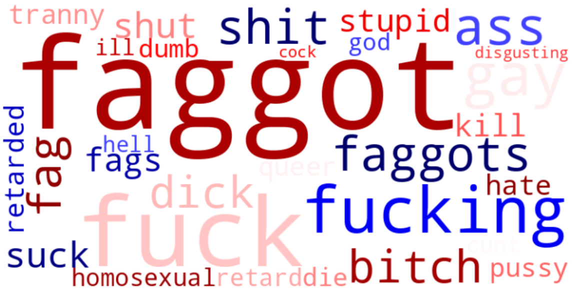


Figure 4: Word cloud of top most frequent words of anti-LGBTQIA+ user-generated texts.

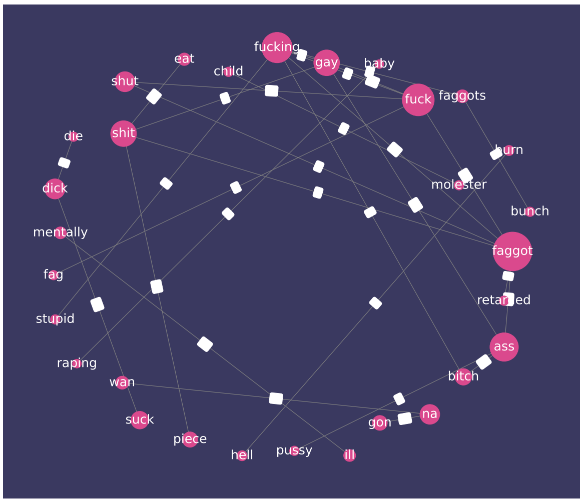


Figure 5: Bigram network graph of strongest associations between commonly occurring offensive contents of anti-LGBTQIA+ user-generated texts.

Table 1 and Figure 5 present the top 30 strongest associations between commonly occurring offensive contents in anti-LGBTQIA+ user-generated texts. The most frequent bigram, “fucking faggot,” appears 33 times, followed by other highly offensive phrases such as “suck dick” (24 times) and “fuck faggot” (23 times). Many bigrams include slurs targeting LGBTQIA+ individuals (e.g., “shut faggot," "gay shit," "retarded faggot") and general profanity combined with aggression (e.g., "burn hell," "dick die," "baby raping").

5.2 Model Performance

The results presented in Table 2 and Figure 3 highlight variations in performance among different LLMs in detecting anti-LGBTQIA+ user-generated texts. Notably, DeepSeek-V3 proves to be the best-performing model. It achieves the highest Acc and TPR of 0.866, along with the highest F1 of 0.871. Furthermore, it maintains the lowest FPR, indicating its high predictive performance in detecting anti-LGBTQIA+ user-generated text.

Next, Llama3.3-70B is the second-best method, achieving the highest Prec of 0.891, which underscores its effectiveness in minimizing false positives. It also shows notable high performance with Acc, TPR, F1, and FPR, which are just ranked below DeepSeek-V3.

Although the performance on all metrics remains comparatively lower than DeepSeek-V3 and Llama3.3-70B, GPT-4o-mini and GPT-4o have the best calibrated probabilistic predictions achieving the lowest $Brier$, with values of 0.551 and 0.585, respectively. Regarding GPT-o1-mini, it underperforms across all evaluation metrics compared to other LLMs, suggesting limitations in its effectiveness for the anti-LGBTQIA+ user-generated texts classification task.

Importantly, the $Brier$ values of all LLMs are notably high with all above 0.5, suggesting the necessity for improving probability calibration across them despite some delivering lower scores than others. Generally, LLMs exhibit overconfidence (Yu et al., 2024), as demonstrated by their calibration curves (Figure 3) falling below the 45° perfect calibration line (Bol et al., 2012). This suggests that the predicted probabilities (confidence score c as explained in Section 3.2) are higher than the actual likelihood of respective outcomes.

5.3 Error Analysis of Misclassified Texts

DeepSeek-V3 is proven to be the best-performing model in the previous section, but it still has limitations in accurately classifying anti-LGBTQIA+ user-generated texts. A closer examination of misclassified words reveals words that contribute to these errors (see Figures 6, 7, and Table 3).

High-frequency identity-related terms ("gay," "trans," "lesbian," "LGBT") frequently co-occur with neutral and offensive words, indicating the model's difficulty in distinguishing between discussions and harmful rhetoric. Additionally, offensive terms such as "fuck," "bitch," and "faggot" form toxic associations ("gay, shit"), underscoring the challenge of separating explicit hate speech from informal language. Furthermore, neutral words like "community," "gender," and "pride" appear in controversial contexts ("gay, community"), revealing limitations in contextual understanding.

These misclassifications highlight underlying sociocultural biases and detection limitations inherent within the model, with terms like "white," "straight," "god," and "fact" often reflecting ideological framing ("straight, white" and "gay, agenda"). Bigrams such as "fuck, stupid" and "gay, marriage" further highlight the model's struggle with contextual nuance, emphasising the need for improved context-aware learning to improve its performance.

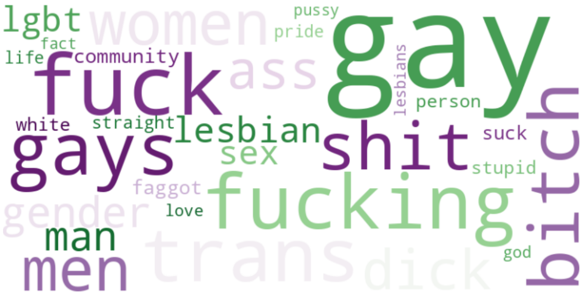


Figure 6: Misclassified samples of anti-LGBTQIA+ user-generated texts from DeepSeek-V3 - Word cloud of top most frequent words.

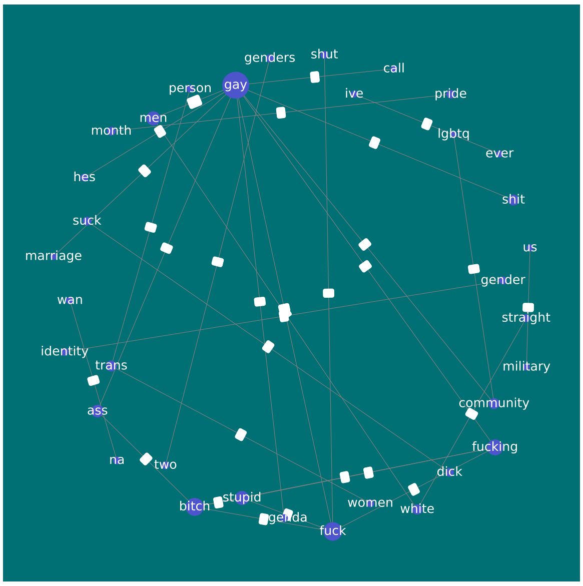


Figure 7: Misclassified samples of anti-LGBTQIA+ user-generated texts from DeepSeek-V3 - Bigram network graph of strongest associations.

Method	Accuracy ↑	TPR ↑	Precision ↑	F1 ↑	FPR ↓	Brier ↓
DeepSeek-V3 (Liu et al., 2024)	**0.866**	**0.866**	*0.890*	**0.871**	**0.160**	0.610
GPT-4o (Hurst et al., 2024)	0.820	0.820	0.878	0.828	0.242	*0.585*
GPT-4o-mini (Hurst et al., 2024)	0.838	0.838	0.880	0.844	0.208	**0.551**
GPT-o1-mini (Aaron et al., 2024)	0.769	0.769	0.832	0.779	0.283	0.610
Llama3.3-70B (Jonas et al., 2025)	*0.863*	*0.863*	**0.891**	*0.868*	*0.168*	0.634

Table 2: Performance comparison of different Large Language Models (LLMs) for detecting anti-LGBTQIA+ user-generated texts. **Bold value**: Best metric. *Italic value*: Second-best metric.

Word Frequency			Association Frequency	
Word	**Percentage**	**Count**	**Association**	**Count**
gay	3.473	224	(’gay’, ’men’)	10
fuck	0.930	60	(’gay’, ’shit’)	8
fucking	0.760	49	(’ass’, ’bitch’)	6
trans	0.744	48	(’fucking’, ’bitch’)	5
gays	0.713	46	(’suck’, ’dick’)	5
bitch	0.651	42	(’gay’, ’community’)	5
shit	0.589	38	(’fuck’, ’stupid’)	5
women	0.496	32	(’fucking’, ’gay’)	5
men	0.496	32	(’two’, ’genders’)	4
dick	0.403	26	(’bitch’, ’fuck’)	4
ass	0.372	24	(’shut’, ’fuck’)	4
gender	0.372	24	(’trans’, ’person’)	4
man	0.341	22	(’gender’, ’identity’)	4
lesbian	0.326	21	(’wan’, ’na’)	4
lgbt	0.310	20	(’pride’, ’month’)	4
sex	0.310	20	(’stupid’, ’fucking’)	4
community	0.310	20	(’hes’, ’gay’)	4
faggot	0.310	20	(’gay’, ’ass’)	4
suck	0.295	19	(’gay’, ’agenda’)	4
person	0.279	18	(’gay’, ’fuck’)	3
stupid	0.279	18	(’fuck’, ’fucking’)	3
white	0.279	18	(’trans’, ’women’)	3
straight	0.264	17	(’stupid’, ’bitch’)	3
life	0.248	16	(’straight’, ’white’)	3
pride	0.248	16	(’lgbtq’, ’community’)	3
love	0.233	15	(’white’, ’men’)	3
god	0.233	15	(’us’, ’military’)	3
pussy	0.217	14	(’call’, ’gay’)	3
lesbians	0.217	14	(’ive’, ’ever’)	3
fact	0.217	14	(’gay’, ’marriage’)	3

Table 3: Misclassified samples of anti-LGBTQIA+ user-generated texts from DeepSeek-V3 - Top 30 most frequent words and strongest bigram associations.

6 Conclusions and Discussions

This research proves the potential of LLMs for real-world applications in identifying anti-LGBTQIA+ user-generated content and underscores the valuable insights that text analysis can provide for topic modelling. These findings play a crucial role in fostering safer digital environments, ultimately improving protections for LGBTQIA+ individuals including their mental health and well-being.

To begin with, regarding the RQs outlined in Section 1, about **RQ1**, our analysis of anti-LGBTQIA+ user-generated texts (see Section 5.1) reveals a high prevalence of derogatory language, hate speech, and aggressive expressions. This can significantly contribute to topic modelling research. These findings underscore the urgent need for effective moderation strategies and improved detection models to mitigate harmful content and foster a safer online environment, improving the mental health and well-being of LGBTQIA+ individuals. Moreover, the proposed framework with the workflow in Section 3.2, including the developed prompt and experiments establish a general pipeline for leveraging LLMs in detecting anti-LGBTQIA+ user-generated texts, addressing **RQ2**.

For **RQ3**, as detailed in Section 5.2 while LLMs demonstrate promising performance in detecting anti-LGBTQIA+ user-generated texts, improvements are still necessary for real-world deployment. Firstly, performance varies across different metrics. DeepSeek-V3 and Llama3.3-70B emerge as the top-performing models; however, their calibration is not as good as GPT-4o and GPT-4o-mini. In contrast, GPT-o1-mini consistently underperforms across all metrics, underscoring its limitations in this task. Notably, despite achieving the highest performance, DeepSeek-V3 and Llama3.3-70B still fall short, with all key metrics (Acc, TPR, Prec, and F1) remaining below 0.9. This highlights the limitations of these LLMs in a zero-shot setting, emphasizing the need for fine-tuning and further development to enhance their reliability and applicability. On top of that, all LLMs exhibit **a notable calibration issue**, tending to be overconfident in their predictions. This overconfidence can lead to increased false positives and false negatives, resulting in unreliable moderation/classification of anti-LGBTQIA+ content. Additionally, it may amplify biases, reduce trust in AI-driven moderation systems, and create challenges in human-AI collaboration by insufficient moderators. Furthermore, as analysed in Section 5.3, although achieving the best-performing model, DeepSeek-V3 has limitations in distinguishing between neutral discussions and harmful rhetoric, struggles with contextual nuance, and exhibits sociocultural detecting limitations in detecting anti-LGBTQIA+ user-generated texts.

The findings of this study establish a strong foundation for future research. Future work should aim to enhance model performance through strategies such as few-shot prompting (Pengyue et al., 2024; Tom et al., 2020) which may significantly improve the predictive capabilities of LLMs in detecting anti-LGBTQIA+ user-generated texts. Additionally, utilising larger-scale datasets with different languages is a crucial next step. Additionally, ensuring demographic representation is critical for assessing LLMs' performance, and fairness across gender, nationality, LGBTQIA+ subgroups, and so on. These advancements will contribute to developing a robust, fair, and generalisable LLM-based anti-LGBTQIA+ user-generated text detection framework for protecting people of the LGBTQIA+ community.

References

Jaech Aaron et al. 2024. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.

Roberto L Abreu and Maureen C Kenny. 2018. Cyberbullying and LGBTQ youth: A systematic literature review and recommendations for prevention and intervention. *Journal of Child & Adolescent Trauma*, 11:81–97.

Carolina Are, Catherine Talbot, and Pam Briggs. 2024. Social media affordances of LGBTQIA+ expression and community formation. *Convergence*, page 13548565241296628.

Adwita Arora, Aaryan Mattoo, Divya Chaudhary, Ian Gorton, and Bijendra Kumar. 2024. MEnTr@ LT-EDI-2024: Multilingual ensemble of transformer models for homophobia/transphobia detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 259–264.

Bol, Linda, Hacker, and Douglas J. 2012. *Calibration*, pages 495–498. Springer US, Boston, MA.

Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten W Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and hate speech: Structured explanations make social media moderators faster. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Waterschoot Cedric et al. 2022. Detecting minority arguments for mutual understanding: a moderation tool for the online climate change debate. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6715–6725.

Han Chi et al. 2024. LM-Infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 62nd Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3991–4008.

Mirko Franco, Ombretta Gaggi, and Claudio E Palazzi. 2024. Integrating content moderation systems with large language models. *ACM Transactions on the Web*.

Gianluca Gini and Dorothy L Espelage. 2014. Peer victimization, cyberbullying, and suicide risk in children and adolescents. *Jama*, 312(5):545–546.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Aric Hagberg and Drew Conway. 2020. Networkx: Network analysis with Python. *URL: https://networkx. github. io*.

Steven A Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A Riegler, Pål Halvorsen, and Sravanthi Parasa. 2022. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979.

Melissa K Holt, Alana M Vivolo-Kantor, Joshua R Polanin, Kristin M Holland, Sarah DeGue, Jennifer L Matjasko, Misty Wolfe, and Gerald Reid. 2015. Bullying and suicidal ideation and behaviors: A meta-analysis. *Pediatrics*, 135(2):e496–e509.

Jun Sung Hong, Raúl Navarro, and Michelle F Wright. 2025. Adolescent cyberbullying: A worldwide concern. In *Encyclopedia of Information Science and Technology, Sixth Edition*, pages 1–22. IGI Global.

Danqing Hu, Bing Liu, Xiaofeng Zhu, Xudong Lu, and Nan Wu. 2024. Zero-shot information extraction from radiological reports using ChatGPT. *International Journal of Medical Informatics*, 183:105321.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Cho Hyundong et al. 2024. Can language model moderators improve the health of online discourse? In *Proceedings of the 62nd Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7471–7489.

Cascalheira Cory J et al. 2024. The LGBTQ+ minority stress on social media (missom) dataset: A labelled dataset for natural language processing and machine learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1888–1899.

Yuping Jin. 2017. Development of word cloud generator software based on python. *Procedia engineering*, 174:788–792.

Wihl Jonas et al. 2025. Data extraction from free-text stroke ct reports using GPT-4o and Llama-3.3-70B: The impact of annotation guidelines. *medRxiv*, pages 2025–01.

Taylor Jordan et al. 2024. Cruising Queer HCI on the DL: A Literature Review of LGBTQ+ People in HCI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Yubo Kou and Xinning Gui. 2020. Mediating community-AI interaction through situated explanation: the case of AI-Led moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27.

Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2024. Self-prompting large language models for zero-shot open-domain QA. In *Proceedings of the 62nd Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 296–310.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Meta. 2025. Llama API platform. Accessed: 19 February 2025.

Falk Neele et al. 2024. Moderation in the wild: Investigating user-driven moderation in online discussions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 992–1013.

Joel Nothman, Hanmin Qin, and Roman Yurchak. 2018. Stop word lists in free open-source software packages. In *Proceedings of workshop for NLP open source software (NLP-OSS)*, pages 7–12.

L Haimson Oliver et al. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–35.

OpenAI. 2025. OpenAI API platform. Accessed: 19 February 2025.

Parth Patel. 2025. Detection of hate speech against lgbt+ on social media. Accessed: 2025-03-14.

Jia Pengyue et al. 2024. Mill: Mutual verification with large language models for zero-shot query expansion. In *Proceedings of the 62nd Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2498–2518.

Sachdeva Pratik et al. 2022. The measuring hate speech corpus: Leveraging Rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.

Jha Prince et al. 2024. Memeguard: An LLM and VLM-based framework for advancing content moderation via meme intervention. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers).*

O'Hagan Ross et al. 2023. The accuracy and appropriateness of ChatGPT responses on nonmelanoma skin cancer information using zero-shot chain of thought prompting. *JMIR dermatology*, 6:e49889.

Kaspar Rufibach. 2010. Use of brier score to assess binary predictions. *Journal of clinical epidemiology*, 63(8):938–939.

Masud Sarah et al. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863.

Ryan Schey and Stephanie Anne Shelton. 2023. Queer (ing) and trans (ing) critical media literacies in response to Anti-LGBTQIA+legislation and policies. *The International Journal of Critical Media Literacy*, 3(2):73–87.

Markov Todor et al. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

Brown Tom et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Raj Vivek et al. 2024. Conbert-rl: A policy-driven deep reinforcement learning based approach for detecting homophobia and transphobia in low-resource languages. *Natural Language Processing Journal*, 6:100040.

Qiao Wei et al. 2024. Scaling up llm reviews for google ads content moderation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1174–1175.

Yuan Youliang et al. 2024. Does chatgpt know that it does not know? Evaluating the black-box calibration of ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5191–5201.

Yuan Yu et al. 2024. Do llms overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200.