Moderating Harm: Benchmarking Large Language Models for Cyberbullying Detection in YouTube Comments

Amel Muminovic
International Balkan University
Skopje, North Macedonia

ABSTRACT

As online platforms grow, comment sections increasingly host harassment that undermines user experience and well-being. This study benchmarks three state-of-the-art large language models: OpenAI GPT-4.1, Google Gemini 1.5 Pro, and Anthropic Claude 3 Opus, on a corpus of 5 080 YouTube comments drawn from high-abuse videos in gaming, lifestyle, food vlog, and music channels. The dataset comprises 1 334 harmful and 3 746 non-harmful messages in English, Arabic, and Indonesian, annotated independently by two reviewers with almost perfect agreement (Cohen's $\kappa = 0.83$). Each model is evaluated in a strict zero-shot setting with an identical minimal prompt and deterministic decoding, giving a fair multi-language comparison without task-specific tuning. GPT-4.1 achieves the best balance with an F1 score of 0.863, precision of 0.887, and recall of 0.841. Gemini flags the most harmful posts (recall = 0.875) but its precision falls to 0.767 because of frequent false positives. Claude attains the highest precision at 0.920 and the lowest false-positive rate of 0.022, yet its recall drops to 0.720. Qualitative analysis shows that all three models struggle with sarcasm, coded insults, and mixed-language slang. The findings highlight the need for moderation pipelines that combine complementary models, incorporate conversational context, and fine-tune for under-represented languages and implicit abuse. A de-identified version of the dataset, along with the prompts and model outputs, has been made available to support reproducibility and further progress in automated content moderation.

General Terms

Large Language Models, Cyberbullying, Social Media

Keywords

Artificial intelligence, Cyberbullying, Hate Speech, Large Language Models, Natural Language Processing, Social Media

1. INTRODUCTION

Online platforms now mediate much of everyday communication, with more than 4.95 billion active social media accounts worldwide in 2023 [1]. That ubiquity carries risk: recent studies report that between 6.5% and 35.4% of adolescents in the U.S. and Europe have experienced cyberbullying [2], and longitudinal studies link such exposure to elevated anxiety, depression, and suicidal ideation [3]. As the scale and psychological impact of online abuse grow, so does the urgency of developing reliable, context-aware moderation methods, especially for platforms popular with adolescents such as YouTube and TikTok.

1.1 Background

Cyberbullying has become a major digital safety concern because abusive material is persistent, anonymous, and can spread instantaneously across networks [4]. As global connectivity and social media adoption have expanded, online harassment has become more visible, with increasing concern about its scale and severity [5].
Comment sections on social networks, video sharing sites, and forums are particular hotspots. In a 2024 U.S. survey of parents, 79% said their children had encountered cyberbullying on YouTube [6]. A cross-national analysis of 180 000 adolescents in 42 countries also linked problematic social-media use to both cyberbullying victimization and perpetration, with stronger effects among girls [7].
In response, platforms increasingly deploy artificial-intelligence (AI) systems to moderate user-generated content at scale [8, 9]. Understanding where current large language models (LLMs) succeed and fail is therefore critical for building safer online spaces.

1.2 Motivation and Problem Statement

AI systems are now central to content moderation, especially on large platforms where the volume of user-generated posts exceeds human capacity [10]. Yet despite advances in natural language processing, enforcement remains inconsistent. Abusive comments that rely on sarcasm, coded language, or emotional manipulation often slip through AI moderation, even when they clearly violate platform policies. These subtleties continue to pose challenges for automated systems, which often misread tone or miss indirect expressions of harm [11].
This is especially dangerous in cyberbullying, where the impact of a single overlooked comment can be deeply personal [12]. Unlike hate speech or spam, bullying is often subtle: expressed through mockery, group pressure, or repeated jabs that may seem harmless in isolation but accumulate over time. Many academic benchmarks fail to capture this complexity, relying on synthetic or crowd-sourced datasets with clear-cut abuse [13, 14]. Even widely used corpora such as OLID [15] and HateCheck [16] reveal the same

gap: they consist of short, isolated text snippets and miss the multi-turn, emotionally layered exchanges common in YouTube threads. This study addresses that gap by evaluating three most advanced LLMs, GPT-4, Gemini, and Claude, on authentic YouTube comments from videos with documented cyberbullying. This study examines not just whether models detect harmful language, but also how well they handle ambiguity, cultural context, and tone, with a focus on nuanced, emotionally loaded content.

1.3 Objectives of the Study

Four specific objectives were defined:

(1) Benchmark performance - Compare OpenAI GPT-4.1, Google Gemini 1.5 Pro, and Anthropic Claude 3 Opus in detecting cyberbullying within real YouTube data.
(2) Quantify error patterns - Measure false positive and false negative rates and identify recurring misclassifications themes.
(3) Conduct qualitative error analysis - Manually inspect model outputs to assess contextual understanding and pinpoint linguistic edge cases.
(4) Assess robustness across nuance - Evaluate how well each model handles abusive content that is linguistically ambiguous, culturally specific, or indirectly expressed.

2. RELATED WORK

2.1 AI Content Moderation and Toxic Language Detection

Automated content moderation has evolved from simple keyword-based filters to machine learning classifiers and, more recently, transformer-based models. Early systems relied on blacklists or logistic regression trained on hand-labeled corpora, but struggled with negation, sarcasm, and informal language. The advent of BERT [17], RoBERTa [18], and their derivatives enabled context-aware classification by capturing bidirectional dependencies and subtle phrasing nuances. These models improved performance on public benchmarks such as the Jigsaw Toxic Comment dataset and HateXplain [19], driving adoption in industry moderation pipelines. Mathew *et al.* [19] fine-tune BERT-base on HateXplain and report a macro-F_1 of 0.674; a side-by-side comparison with this papers zero-shot LLM scores appears later in Table 4.

However, most benchmarks are composed of isolated sentences and short comments labeled through crowdwork. They rarely reflect the conversational, multi-turn, or emotionally charged dynamics seen in real-world platforms. Furthermore, these datasets often over-represent overt toxicity while under-representing gray area content such as mockery, dismissive sarcasm, or personal digs. As a result, models trained on them perform well in benchmark settings but struggle in live deployments where language is more ambiguous and emotionally expressive [20, 21].

2.2 Challenges in Detecting Implicit and Nuanced Abuse

Detecting implicit harm remains one of the most difficult challenges in AI-based moderation. Sarcasm, coded language, cultural slang, and subtextual insults frequently evade model detection, even in LLMs. For example, a phrase like “She’s always such a queen, right?” may be an insult depending on tone and context, but appears harmless in isolation. Studies in sarcasm detection [22], adversarial misspellings, irony detection, and tone modeling have attempted to address this gap, but progress remains limited [23].

Recent research also suggests that models often over-flag emotionally intense or sensitive language even when used in supportive contexts, leading to user frustration and moderation fatigue. This tension between under-flagging subtle harm and over-flagging emotional expression makes fine-tuned moderation especially difficult. Studies have proposed integrating tone analysis, conversational history [24], and speaker intent as ways to improve subtle abuse detection, but these are rarely deployed at scale.

2.3 Moderating Multilingual and Under-resourced Content

Although English dominates most training corpora, abuse occurs in every language, and much of it is multilingual, transliterated, or mixed with emojis and slang [25]. Studies consistently show significant drops in model accuracy when applied to languages such as Arabic [26] and Hindi [27], particularly when users employ spelling variation or obfuscation to evade filters [28]. Code-switching and transliteration further degrade detection performance, as models frequently miss common abusive patterns or misinterpret them out of context [29, 30].

Cross-lingual models and prompt-tuned systems have demonstrated moderate performance improvements, especially when fine-tuned on translated or augmented datasets [31]. However, significant language imbalance remains, particularly in online spaces with small yet highly active non-English user communities [32]. For platforms with global audiences, this poses a critical equity issue: harmful content in English is far more likely to be effectively moderated than equivalent abuse in low-resource languages or regional dialects [33].

2.4 Evaluating LLMs for Safety and Fairness in Moderation

The rise of general-purpose LLMs such as GPT-4, Claude, and Gemini has driven increasing interest in their application to content moderation tasks. Recent research indicates that these models can outperform traditional task-specific classifiers in zero-shot or few-shot scenarios, especially on previously unseen or nuanced content [34]. Their ability to dynamically incorporate platform policies, contextual reasoning, or nuanced moderation guidelines directly into their prompting strategies enhances their flexibility and adaptability [35].

However, the performance of LLMs in moderation tasks is highly sensitive to prompt phrasing, target domains, and cultural nuances [36, 37]. Sociocultural audit frameworks have recently been proposed, employing persona-based prompts [38] and synthetic demographic simulations to systematically evaluate model fairness and reveal hidden biases or blind spots [39, 40]. These frameworks highlight how LLM responses may shift significantly based on perceived user identity or topical framing, raising substantial concerns around fairness, consistency, and equitable moderation [41].

While LLMs offer substantial advantages in scalability, adaptability, and reduced reliance on explicit rule-based configurations [42], their inherent dependence on pretraining data, which often contain historical biases, and the opacity of their decision-making processes continue to present barriers to safe and transparent deployment in moderation workflows [43].

3. METHODOLOGY

3.1 Case Selection and Data Collection

Comments were retrieved via the YouTube Data API between 15 April 2024 and 15 May 2025.

The final corpus covers four public videos from distinct creators and content domains such as gaming, lifestyle, food vlog, and music. To verify that each video was indeed a cyber-bullying hotspot, a pilot crawl of 1 000 comments per candidate clip was first run, retaining only those whose pilot abuse rate exceeded 20%. From the four retained videos, a uniform random sample was then drawn, yielding 5 080 comments in total. Only public endpoints were accessed; private, removed, or shadow-banned comments are not included.

3.2 Data Cleaning and Anonymization

Each comment underwent a two–step preprocessing pipeline. First, text was normalized (UTF-8 decoding, whitespace trimming, emoji and punctuation were preserved). Second, personally identifiable information including user names, real names, phone numbers, e-mail addresses, links, and explicit geo-markers was either deleted or replaced by neutral placeholders such as "UserNameProtected". No other linguistic content was altered. This procedure leaves the semantic core of each message intact while mitigating re-identification risk, allowing the corpus to be shared for research without exposing private data.

3.3 Data Labeling and Ground Truth

All 5 080 comments were independently annotated by two reviewers (the first author and a second annotator holding a B.Sc. in Software Engineering) using a binary rubric:

3.3.1 Harmful. Bullying, sustained harassment, severe personal insults, or death-related threats

3.3.2 Not harmful. Neutral, supportive, off-topic, or otherwise non-abusive remarks

Inter-rater reliability was high, with 91.0% raw agreement and $\kappa = 0.83$, which is considered almost perfect.[1] Disagreements (9% of instances) were adjudicated by discussion until consensus, and that label was written to the final ground truth file. Labels were never exposed to the models.

Non-English comments, including those in Arabic and Indonesian, were labeled using a combination of manual translation, consultation with fluent speakers, and, when necessary, language translation tools. While consistency was a priority, minor errors in interpreting slang or cultural references are possible and are noted as a study limitation.

3.4 Evaluated Models

Three widely used large language models were selected for evaluation:

(1) OpenAI GPT-4.1

(2) Google Gemini 1.5 Pro

(3) Anthropic Claude 3 Opus (2024-02-29)

Each model was tasked with classifying the same 5 080 user-generated comments using identical prompt structures to ensure fairness and consistency across systems. These models were selected due to their wide adoption, advanced reasoning capabilities, and public accessibility through API services, making them representative of current best practices in AI content moderation.

3.5 Prompt Design and Justification

All three models were queried with the same minimal prompt, using `temperature = 0.0` to remove randomness and `max_tokens = 1` to ensure the model outputs only a single token, allowing for a strict binary label (0 or 1). Locking these parameters serves three goals:

(1) Fair cross-model comparison. One fixed prompt eliminates tuning bias and attributes performance differences solely to the models.

(2) Deterministic, reproducible outputs. With temperature 0.0 and a short token cap, each comment always returns the same "0" or "1," so any researcher can replicate the results with a single API call.

(3) Scope control. The study provides a baseline snapshot of zero-shot LLM moderation on real-world data rather than an optimized prompt. Systematic prompt engineering is left to future work.

The fixed prompt used in every call is shown below.

```
Analyze the following comment and
determine if it contains cyberbullying,
harassment, or harmful language. Make
sure to understand the context, tone and
potential impact. Respond with 0 if it
is not harmful and 1 if it is harmful.
Comment is: <COMMENT_TEXT>
```

The prompt offers no label definitions beyond the binary mapping, requiring each model to infer its own policy from plain language while still accounting for nuance.

3.6 Evaluation Metrics

Each model's output was compared against the manually labeled ground truth using standard classification metrics:

—True Positives (TP): Model correctly flagged a harmful comment

—True Negatives (TN): Model correctly identified a non-harmful comment

—False Positives (FP): Model incorrectly flagged a non-harmful comment

—False Negatives (FN): Model failed to flag a harmful comment

From this, the following performance metrics were calculated:

Precision measures the proportion of comments that the model correctly identified as harmful out of all the comments it flagged as harmful. In other words, it reflects how accurate the model is when it predicts that a comment is harmful. It is calculated using Equation (1).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Recall looks at all the harmful comments in the dataset and measures how many the model correctly identified. A higher recall

[1] Landis and Koch, 1977.

Table 1. Dataset Distribution

Class	Count	Percentage
Harmful	1 334	26.3%
Non-harmful	3 746	73.7%
Total	5 080	100%

means the model caught more of the actual harmful content. The corresponding formula is shown in Equation (2).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

F1 Score provides a balance between precision and recall. It is especially useful when both false positives and false negatives carry equal importance. The formula is defined in Equation (3).

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Accuracy shows the overall percentage of correct predictions, covering both harmful and non-harmful cases. While it gives a general sense of performance, it may be less informative in imbalanced datasets. The formula appears in Equation (4).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

Together, these metrics offer a well-rounded view of the models' moderation performance across different types of classification errors.

Macro-averaged (class-balanced) scores were also calculated with the 26% / 74% split and yielded the same model ranking, so the detailed numbers are omitted for brevity.

3.7 Manual Review and Model Comparison

Quantitative scores alone do not reveal *why* a system succeeds or fails, so a post-hoc qualitative review was carried out. Two annotators independently inspected a stratified sample of 200 disagreements, consisting of 50 false positives, 50 false negatives, and 100 edge case ties, drawn in equal proportion from all three models. For each comment, the review considered (i) whether the human ground truth should stand and (ii) what linguistic cues might have misled the model (e.g., sarcasm, coded slurs, emoji, topic drift). Notes from this exercise were grouped into recurring error themes that inform the Discussion section. The procedure does not alter any numeric results but clarifies how each system handles nuance, context, and cultural references.

4. RESULTS

4.1 Dataset Composition

The evaluation dataset consisted of 5 080 comments, including 1 334 harmful and 3 746 non-harmful instances. Harmful content made up about 26.3% of the total dataset, indicating a moderate class imbalance that could influence model metrics like precision and recall. Table 1 provides the class breakdown.

4.2 Model-Level Classification Performance

Each model was evaluated using raw classification outcomes and derived performance metrics. Table 2 presents the counts for true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each model.

All three models showed strong ability to correctly identify non-harmful comments, with true negative counts above 3 300. Gemini identified the highest number of harmful comments, with 1 167 true positives, but also had the highest number of false positives, mistakenly labeling 354 non-harmful comments as harmful. GPT had a solid performance overall, producing 1 122 true positives and only 143 false positives. Claude was more selective, correctly identifying 961 harmful comments while minimizing false positives to 83. From these outcomes, standard classification metrics were calculated: accuracy, precision, recall, F1 score, and false positive rate. These provide a clearer picture of each model's behavior in detecting harmful content. Table 3 displays the results.

Table 2. Prediction Counts for Each Model

Model	TP	TN	FP	FN
GPT	1 122	3 603	143	212
Gemini	1 167	3 392	354	167
Claude	961	3 663	83	373

Table 3. Evaluation Metrics Based on Prediction Outcomes

Model	Accuracy	Precision	Recall	F1 Score	FPR
GPT	0.930	0.887	0.841	0.863	0.038
Gemini	0.897	0.767	0.875	0.818	0.095
Claude	0.910	0.920	0.720	0.808	0.022

Table 4. Macro-F_1 for a classic fine-tuned baseline versus zero-shot LLMs

Model	Dataset / Setting	Macro-F_1
BERT-base (fine-tuned) [19]	HateXplain	0.674
GPT-4.1 (zero-shot, *this work*)	YouTube 5 080	0.863
Gemini 1.5 Pro (zero-shot, *this work*)	YouTube 5 080	0.818
Claude 3 Opus (zero-shot, *this work*)	YouTube 5 080	0.808

GPT achieved the most balanced performance across the board. It reached an F1 score of 0.863, with high values for both precision at 0.887 and recall at 0.841. Claude stood out for its high precision, achieving 0.920, and maintained the lowest false positive rate at 0.022. However, this conservative stance came at the cost of recall, which was limited to 0.72. Gemini prioritized identifying as many harmful comments as possible, leading to the highest recall of 0.875, but its precision dropped to 0.767 and its false positive rate increased to 0.095. Representative misclassified comments are discussed in Section 5.1.

4.3 Baseline comparison with a fine-tuned transformer

Mathew *et al.* [19] fine-tune BERT-base on the HateXplain corpus and report a macro-F_1 of 0.674 (Table 6 of their paper). Table 4 sets that published baseline beside the zero-shot scores obtained here. Because the corpora differ, these figures are not a head-to-head benchmark; the BERT row is included only as a widely cited transformer reference point.

4.4 Observations and Comparisons

The models displayed clear differences in their handling of harmful content. GPT offered a strong balance, combining effective detection with a relatively low error rate. This made it suitable for environments where both safety and user experience matter. Claude's high precision and low false positive rate reflect a more cautious approach, making it better for contexts where false accusations must be minimized. Gemini, with the highest recall, is more aggressive, potentially fitting platforms that prioritize safety even at the cost of occasional over-flagging.

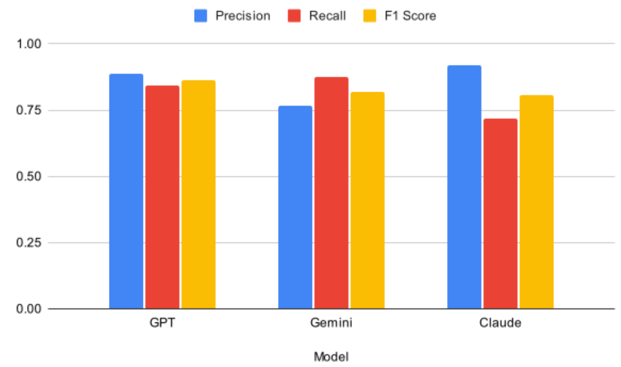

Fig. 1. Comparison of precision, recall, and F1 score across evaluated models

Each model reflects a distinct moderation philosophy. GPT is consistent and balanced, Claude is precise and conservative, and Gemini is proactive and broad-reaching. The choice of model should depend on the platform's goals and tolerance for errors in either direction.

4.5 Visual Comparison

Figure 1 compares the models using precision, recall, and F1 score. These metrics highlight how each model prioritizes different aspects of content moderation.
The visual comparison supports earlier findings. GPT maintained a steady balance between precision and recall. Claude achieved top precision but lagged in recall. Gemini led in recall, though with lower precision. These patterns underline that no single model is best across all metrics. Instead, the optimal choice depends on what kind of moderation a platform values most. A platform that can tolerate more false positives may prefer Gemini, while one focused on minimizing moderation mistakes may lean toward Claude. In higher-risk environments, a combination of models could help match the tone, topic, and user profile of each comment more effectively.

4.6 Model Agreement Analysis

To better understand the consistency of predictions across models, this study analyzed agreement between GPT-4.1, Google Gemini 1.5 Pro, and Anthropic Claude 3. Out of the 5 080 moderated comments, all three models predicted the same label for 4 255 of them, which corresponds to 83.76% of the dataset. In 4 170 of those cases (82.09% of total), the shared prediction matched the human annotation.
In the remaining 825 comments (16.24%), the models disagreed with each other. These disagreement cases often involved sarcasm, emotionally charged phrases, or borderline language. Such examples are frequently misclassified by one or more models in isolation. The results suggest that combining models in an ensemble moderation setup could improve reliability in detecting complex or ambiguous content. A detailed breakdown of these agreement patterns is provided in Table 5.

Table 5. Summary of model agreement and correctness across all 5 080 comments

Metric	Count	Percentage
Total comments	5 080	100%
Full agreement (all 3 models predicted the same)	4 255	83.76%
Full agreement and correct	4 170	82.09%
Disagreement cases	825	16.24%

5. DISCUSSION

5.1 False Positives and Surface-Level Flagging

To better understand model over-flagging, an analysis was conducted on cases where AI models marked comments as harmful while human reviewers did not. These false positives were grouped by model agreement patterns to highlight shared tendencies and specific weaknesses. While most flagged content reflected strong language or emotional tone, it often lacked any abusive intent. Four common patterns were observed, as outlined below.

5.1.1 All Three Models vs. Human: Shared Over-flagging Patterns. In a small number of cases, only 12 out of 5 080, all three models (GPT, Claude, and Gemini) identified comments as harmful, while human reviewers did not. These examples typically involved informal phrasing, sarcasm, or emotionally expressive language. For instance, the comment "Y'all are so toxic" was likely flagged due to tone, despite not targeting a specific individual. Another example, "Omg...321lbs! That's so sad," lacked context and may have been flagged merely for mentioning weight. These instances illustrate how all three models tend to rely on surface-level cues, such as slang or emotional keywords, without fully assessing intent or context.

5.1.2 GPT: Over-Flagging Emotional/Mental-Health Language. GPT's false positives often involved emotionally charged or mental health-related phrasing that human reviewers did not consider harmful. These included expressions of concern, frustration, or informal critique that lacked abusive intent. GPT frequently flagged comments referencing emotional breakdowns, support for recovery, or dramatic language about personal change. While this may reflect a cautious stance toward sensitive topics, it also suggests an overreliance on trigger phrases without accounting for context or tone.

5.1.3 Gemini: Over-Flagging Concern and Instability. Gemini frequently flagged comments that referenced mental health, personal change, or emotional concern, even when human reviewers found no clear harm. Many of these comments offered support, expressed worry, or questioned the authenticity of a user's behavior in dramatic or sarcastic terms. Phrases like "please take care of yourself" or "you can't keep scaring everyone like this" were often interpreted by Gemini as harmful. This suggests a tendency to over-flag emotionally sensitive or speculative language, particularly when it touches on perceived instability or trauma.

5.1.4 Claude: Over-Flagging Sarcasm and Hyperbole. Claude's false positives often involved sarcastic remarks, informal roasts, or exaggerated critiques that lacked targeted hostility. Many flagged comments included casual profanity, meme references, or expressive language, for example, jokes about appearance, pop culture, or emotional overreactions. In several cases, Claude appeared to react to tone or strong phrasing rather than the presence of actual harm. This suggests that Claude may place greater weight on civility and politeness, leading it to flag socially edgy but benign content more frequently than human reviewers.

These patterns align with Table 3: Claude's high precision shows caution, Gemini's lower precision reflects aggressiveness, and GPT sits between.

5.2 False Negatives: Sarcasm, Subtext, and Missed Harm

5.2.1 All Models vs. Human: Missed Sarcasm, Mockery, and Coded Harassment. Several harmful comments were missed by all three models, highlighting common blind spots around sarcasm, ridicule, and indirect hostility. Many of these remarks used emojis, mock praise, or exaggerated phrasing to insult, humiliate, or incite collective disdain. Examples included references to "clown emojis," indirect threats like "he feeds off his haters, just ignore him to make him fall off," or comparisons to breakdowns and mental illness framed as jokes.

5.2.2 GPT: Missed Hostility Behind Humor. GPT missed a substantial number of harmful comments flagged by human reviewers due to sarcasm, coded ridicule, or indirect hostility. Phrases like "every time I see you in the hospital I smile" conveyed sustained mockery and dismissiveness toward someone's well-being. These misses suggest that GPT may rely too heavily on surface-level tone and literal phrasing.

5.2.3 Gemini: Missed Harm in Public Shaming. Gemini failed to flag a range of comments often involving mockery or aggression wrapped in sarcasm or cultural shorthand. Some escalated into public shaming, such as calling for mass unsubscribing or questioning the person's sanity. Gemini's tendency to underreact suggests it may struggle with patterns of collective ridicule.

5.2.4 Claude: Missed Sarcasm and Faux Concern. Claude frequently overlooked comments like "Send this man back to Arkham," which ridicule the creator through implication. It also failed to detect hostile rants accusing the creator of deception or instability, showing a reluctance to flag comments unless explicitly hostile.

5.3 Risks of Missed Moderation in Sensitive Cases

False negatives are not equal in impact. While a single missed insult may seem minor, its harm can compound over time, especially in emotionally vulnerable environments. Several missed comments in this study targeted known vulnerabilities, mocked prior hospitalization, or questioned a creator's mental health. These remarks, though subtle or sarcastic, reinforce harmful narratives and may discourage affected users from seeking help or participating in the platform.

More troubling were coordinated suggestions to "unfollow so he falls off" or "ignore him until he cracks again." These comments reflect collective hostility rather than isolated aggression. When such messages go unflagged, they enable mob harassment campaigns that can lead to reputational harm, social withdrawal, or emotional distress for creators. In several cases, these remarks appeared alongside strings of other taunts, suggesting that their impact is not just in their wording but in their cumulative pressure.

This highlights a key limitation of single-comment moderation: by treating each comment in isolation, models miss larger patterns of piling-on or manipulation. Moderation systems must evolve to track conversation history, detect repeated targeting, and recognize coded language that signals coordinated behavior. Especially in cases involving mental health, even a few missed comments can shift the tone of a thread and undermine user safety. Future systems should include tools for temporal tracking, conversational memory, and risk-tiered escalation to reduce the long-term impact of missed moderation.

5.4 Humor, Satire, and Ambiguity in Model Moderation

Another common source of disagreement between human reviewers and LLMs involved ambiguous comments that used humor, satire, or double meaning. These false positives often lacked direct hostility but contained emotionally charged phrasing, exaggerated tone, or social critique. Several examples reveal how models misinterpret stylistic or cultural cues.

One such comment was "Can't sing for nothing!!!!", flagged by GPT despite being a common form of exaggerated critique in entertainment contexts. Similarly, "2:33 too much jiggling" was flagged by Gemini, likely due to its focus on physical attributes, though the intent may have been commentary rather than harassment.

Comments referencing public figures or using layered sarcasm were also commonly misclassified. For instance, "It sounds like UserNameProtected Cobain's remembrance video" was flagged by GPT, Gemini, and Claude, though human reviewers judged it to be a stylistic comparison rather than an attack. Another flagged comment, "You just gonna act like nothing happened?..." shows how models can mistake rhetorical or skeptical questions for aggression, especially when stripped of conversational context.

Claude and Gemini in particular flagged remarks like "UserNameProtected ain't even real" that are commonly used in meme culture or satire. These comments, while edgy, were not interpreted as harmful by human reviewers. However, the models appeared sensitive to tone and phrasing that hinted at ridicule or disbelief.

These cases underscore the limits of current models in interpreting intent, especially when humor blurs the line between critique and cruelty. Without access to conversational history or platform-specific norms, LLMs often make conservative judgments, flagging emotionally charged or socially playful content as harmful.

For moderation systems, this presents a difficult trade-off. Over-flagging benign humor risks alienating users and undermining trust in automated systems, while under-flagging allows veiled insults or passive aggression to persist. Future moderation pipelines may benefit from humor-aware classifiers or community-tuned thresholds that distinguish cultural satire from actual abuse.

6. FUTURE WORK

The open issues are ranked below from highest near-term impact to longer-term research challenges.

(1) Thread-level and multimodal context. Adding preceding turns, video transcripts, or image cues is the fastest way to cut many sarcasm and piling-on errors found in this study.

(2) Ensemble and risk-tiered pipelines. A two-stage approach (first a high-recall filter, then a high-precision check) can be deployed with existing APIs. Measuring latency, cost, and user impact belongs in the next round of experiments.

(3) Implicit and coded abuse corpora. New datasets that label irony, metaphor, and insider slang will let researchers test prompt tweaks and specialised pre-training for covert hostility. This requires a dedicated annotation effort.

(4) User-facing feedback loops. Controlled trials comparing silent removal, warning prompts, and educational pop-ups could clarify which intervention nudges commenters toward civility most effectively.

(5) Long-term mental-health outcomes. Ultimately, evidence is needed to show that better moderation lowers anxiety, self-harm ideation, or community churn. Partnering with mental-health researchers and analyzing de-identified longitudinal data are goals for later phases.

7. CONCLUSION

This paper benchmarks OpenAI GPT-4.1, Google Gemini 1.5 Pro, and Anthropic Claude 3 Opus on 5 080 YouTube comments drawn from high-abuse threads in gaming, lifestyle, food vlog, and music channels. Each model was evaluated with the same prompt under deterministic settings. Gemini identified the largest share of harmful content, achieving recall of 0.875, yet its precision dropped to 0.767 because of frequent false positives. Claude reached precision of 0.920 and the lowest false-positive rate of 0.022, although its recall fell to 0.720. GPT-4.1 delivered the best overall balance with an F1 score of 0.863, precision of 0.887, and recall of 0.841. Qualitative analysis showed that sarcasm, coded insults, and mixed-language slang remain persistent blind spots for all three models. These findings make clear that no single system meets every moderation requirement. Practical pipelines should combine complementary models, incorporate conversational context, and fine-tune for under-represented languages and subtle forms of abuse.

For deployment, platforms can map each model's strengths to specific risk tiers. A site that values maximum coverage can run Gemini as the first-pass filter and then route its flags to Claude or a human queue to reduce false alarms. Low-latency chat services may prefer using GPT alone for a balanced trade-off between recall and precision. Logging each disagreement and feeding it into periodic retraining supports continuous improvement, auditability, and fairness, contributing to AI-safety goals that limit user exposure to harmful content without over-silencing benign speech.

The fully de-identified dataset, prompt text, and model outputs are openly released to support reproducible research and to foster further progress in automated content moderation.

8. LIMITATIONS

This study has several limitations:

(1) Zero-shot prompts: All models were evaluated using identical zero-shot prompts without any model-specific fine-tuning or prompt optimization. This setup maximizes comparability but may understate each model's peak performance.

(2) Comment-level context: Each comment was assessed in isolation, with no access to preceding messages or thread history. As a result, context-dependent abuse, sarcasm, or coordinated behavior across comments may go undetected.

(3) Cultural bias in labeling: Human annotations came from two reviewers sharing similar linguistic and cultural backgrounds, which could bias judgments of sarcasm, slang, or ambiguous phrasing.

(4) Limited language scope: Although a few comments were in Arabic or Indonesian, the dataset is overwhelmingly English. Therefore, the findings of this study should not be generalized to truly multilingual moderation performance.

(5) No BERT-based baseline: Traditional transformer classifiers like BERT and RoBERTa have already been widely studied in toxic language detection tasks [17, 18]. This work focuses instead on production-grade LLMs that are currently deployed or considered for content moderation at scale. Including legacy baselines would provide limited additional insight relative to the study's primary goal of evaluating state-of-the-art zero-shot LLM performance.

9. ETHICAL CONSIDERATIONS

This study uses only publicly available YouTube comments. All user identifiers and any personally identifying details were removed or replaced with neutral placeholders before analysis, so no individual can be re-identified. Because the data are de-identified and pose minimal risk, no formal Institutional Review Board (IRB) review was required.

To address the cultural-bias risk noted in the Limitations section, future releases will include reviews from annotators outside the authors' demographic group and will make the labeling guide public so external researchers can audit or challenge specific decisions.

10. DATA AVAILABILITY

The full, de-identified comment corpus, together with the human labels and model predictions, is openly available at `https://github.com/Ammce/papers/tree/main/llm-cyberbullying-moderation%20`. To minimize re-identification risk, user names, links, and YouTube comment IDs have been removed. The list of source-video IDs can be shared with editors or qualified researchers under a non-disclosure agreement. Exact model prompts and API parameters used in this study are provided in the same repository under `prompts/`.

11. REFERENCES

[1] B. Dean, "Social media usage & growth statistics," *Backlinko*, Feb. 21, 2024. [Online]. Available: https://backlinko.com/social-media-users

[2] A. B. Barragán Martín *et al.*, "Study of cyberbullying among adolescents in recent years: A bibliometric analysis," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, p. 3016, Mar. 2021. doi:10.3390/ijerph18063016

[3] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Arch. Suicide Res.*, vol. 14, no. 3, pp. 206–221, 2010. doi:10.1080/13811118.2010.494133

[4] C. P. Barlett, "Anonymously hurting others online: The effect of anonymity on cyberbullying frequency," *Psychol. Pop. Media Cult.*, vol. 4, no. 2, pp. 70–79, 2015. doi:10.1037/a0034335

[5] L. Huang *et al.*, "The severity of cyberbullying affects bystander intervention among college students: The roles of feelings of responsibility and empathy," *Psychol. Res. Behav. Manag.*, vol. 16, pp. 893–903, Mar. 2023. doi:10.2147/PRBM.S397770

[6] A. Vigderman, "Cyberbullying: Twenty crucial statistics for 2024," *Security.org*, Oct. 9, 2024. [Online]. Available: https://www.security.org/resources/cyberbullying-facts-statistics

[7] W. Craig *et al.*, "Social media use and cyber-bullying: A cross-national analysis of young people in 42 countries," *J. Adolesc. Health*, vol. 66, no. 6, pp. S100–S108, Jun. 2020. doi:10.1016/j.jadohealth.2020.03.006

[8] M. H. Ribeiro, J. Cheng, and R. West, "Automated content moderation increases adherence to community guidelines," in *Proc. ACM Web Conf. (WWW)*, 2023, pp. 2666–2676. doi:10.1145/3543507.3583275

[9] S. Wang and K. J. Kim, "Content moderation on social media: Does it matter who and why moderates hate speech?" *Cyberpsychol. Behav. Soc. Netw.*, vol. 26, no. 7, pp. 527–534, Jul. 2023. doi:10.1089/cyber.2022.0158

[10] T. Gillespie, "Content moderation, AI, and the question of scale," *Big Data Soc.*, vol. 7, no. 2, pp. 1–5, Jul. 2020. doi:10.1177/2053951720943234

[11] H. Lopez and S. Kübler, "Context in abusive language detection: On the interdependence of context and annotation of user comments," *Discourse, Context Media*, vol. 63, Art. no. 100848, Feb. 2025. doi:10.1016/j.dcm.2024.100848

[12] M. van Geel, P. Vedder, and J. Tanilon, "Relationship between peer victimization, cyberbullying, and suicide in children and adolescents: A meta-analysis," *JAMA Pediatr.*, vol. 168, no. 5, pp. 435–442, May 2014. doi:10.1001/jamapediatrics.2013.4143

[13] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*, San Diego, CA, USA, Jun. 2016, pp. 88–93. doi:10.18653/v1/N16-2013

[14] A.-M. Founta *et al.*, "Large scale crowdsourcing and characterization of Twitter abusive behavior," in *Proc. Int. Conf. Web Social Media*, Atlanta, GA, USA, Mar. 2018, pp. 491–500. doi:10.1609/icwsm.v12i1.14991

[15] M. Zampieri *et al.*, "Predicting the type and target of offensive posts in social media," in *Proc. NAACL*, Minneapolis, MN, USA, Jun. 2019, pp. 1415–1420. doi:10.18653/v1/N19-1144

[16] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert, "HateCheck: Functional tests for hate speech detection models," in *Proc. 59th Annu. Meet. Assoc. Comput. Linguistics & 11th Int. Joint Conf. NLP (Long Papers)*, Online, Aug. 2021, pp. 41–58. doi:10.18653/v1/2021.acl-long.4

[17] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423

[18] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, Jul. 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[19] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A benchmark dataset for explainable hate speech detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 17, May 2021, pp. 14867–14875. doi:10.1609/aaai.v35i17.17745

[20] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the worst: Dynamically generated datasets to improve online hate detection," in *Proc. 59th Annu. Meet. Assoc. Comput. Linguistics & 11th Int. Joint Conf. NLP (Long Papers)*, Aug. 2021, pp. 1667–1682. doi:10.18653/v1/2021.acl-long.132

[21] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating neural toxic degeneration in language models," in *Findings Assoc. Comput. Linguistics: EMNLP 2020*, Nov. 2020, pp. 3356–3369. doi:10.18653/v1/2020.findings-emnlp.301

[22] A. Arora, "Sarcasm detection in social media: A review," in *Proc. Int. Conf. Innov. Comput. Commun. (ICICC)*, Dec. 2020, pp. 1–4. doi:10.2139/ssrn.3749018

[23] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, Art. no. 126232, Aug. 2023. doi:10.1016/j.neucom.2023.126232

[24] J. M. Pérez *et al.*, "Assessing the impact of contextual information in hate speech detection," *IEEE Access*, vol. 11, pp. 30575–30590, 2023. doi:10.1109/ACCESS.2023.3258973

[25] A. Muminovic and A. K. Muminovic, "Large Language Models for Toxic Language Detection in Low-Resource Balkan Languages," arXiv preprint arXiv:2506.09992, Jun. 2025. [Online]. Available: https://arxiv.org/abs/2506.09992

[26] H. Mubarak, K. Darwish, and W. Magdy, "Abusive language detection on Arabic social media," in *Proc. 1st Workshop on Abusive Language Online*, Vancouver, Canada, 2017, pp. 52–56. doi:10.18653/v1/W17-3008

[27] T. Mandl *et al.*, "Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages," in *Proc. FIRE*, Kolkata, India, 2019, pp. 14–17. doi:10.1145/3368567.3368584

[28] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan, "'All you need is Love': Evading hate speech detection," in *Proc. 11th ACM Workshop Artif. Intell. Security*, Toronto, Canada, 2018, pp. 2–12. doi:10.1145/3270101.3270103

[29] N. Murikinati, A. Anastasopoulos, and G. Neubig, "Transliteration for cross-lingual morphological inflection," in *Proc. 17th SIGMORPHON Workshop Computational Research Phonetics, Phonology, and Morphology*, Online, Jul. 2020, pp. 189–197. doi:10.18653/v1/2020.sigmorphon-1.22

[30] J. Khanuja, A. Dandapat, A. Srinivasan, S. Sitaram, and M. Choudhury, "GLUECoS: An evaluation benchmark for code-switched NLP," in *Proc. ACL-IJCNLP*, Bangkok, Thailand, 2021, pp. 3575–3585. doi:10.18653/v1/2020.acl-main.329

[31] J. Ranasinghe and M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings," in *Proc. EMNLP*, Online, 2020, pp. 5838–5844. doi:10.18653/v1/2020.emnlp-main.470

[32] Ç. Çöltekin, "A corpus of Turkish offensive language on social media," in *Proc. LREC*, Marseille, France, 2022, pp. 4878–4885. [Online]. Available: https://aclanthology.org/2020.lrec-1.758

[33] E. Pamungkas and V. Patti, "Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon," in *Proc. ACL*, Florence, Italy, 2019, pp. 363–370. doi:10.18653/v1/P19-1051

[34] Y. Liu and M. Zhang, "LLM-Mod: Can Large Language Models Assist Content Moderation?" in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, Rio de Janeiro, Brazil, 2024, pp. 1–12. doi:10.1145/3613905.3650828

[35] F. M. Plaza-Del-Arco, D. Nozza, and D. Hovy, "Respectful or toxic? Using zero-shot learning with language models to detect hate speech," in *Proc. 7th Workshop Online Abuse and Harms (WOAH)*, Singapore, Jan. 2023, pp. 46–52. doi:10.18653/v1/2023.woah-1.6

[36] J. Pavlopoulos *et al.*, "Toxicity detection: Does context really matter?" in *Proc. ACL*, Online, 2020, pp. 4296–4305. doi:10.18653/v1/2020.acl-main.396

[37] A. Baheti, M. Sap, and Y. Tsvetkov, "Just say no: Analyzing the stance of neural dialogue generation in offensive contexts," in *Proc. EMNLP*, Online, 2021, pp. 4846–4859. doi:10.18653/v1/2021.emnlp-main.397

[38] M. Sap *et al.*, “Social bias frames: Reasoning about social and power implications of language,” in *Proc. ACL*, Online, 2020, pp. 5477–5490. doi:10.18653/v1/2020.acl-main.486

[39] I. Solaiman and C. Dennison, “Process for adapting language models to society (PALMS),” Tech. Rep., OpenAI, 2021. [Online]. Available: https://arxiv.org/abs/2106.10328

[40] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings,” in *Proc. NeurIPS*, Barcelona, Spain, 2016, pp. 4356–4364. doi:10.48550/arXiv.1607.06520

[41] H. Welbl, A. Stiennon, and Y. Bai, “Challenges in detoxifying language models,” Tech. Rep., DeepMind, 2021. [Online]. Available: https://arxiv.org/abs/2109.07445

[42] R. Hartvigsen, H. Palangi, and X. He, “Toxigen: Controllable generation of implicit and adversarial toxic text,” in *Proc. ACL*, Dublin, Ireland, 2022, pp. 524–535. doi:10.18653/v1/2022.acl-long.39

[43] E. Bender *et al.*, “On the dangers of stochastic parrots,” in *Proc. FAccT*, Online, 2021, pp. 610–623. doi:10.1145/3442188.344592