Advances in Knowledge-Based Systems, Data Science, and Cybersecurity Received 25-02-2025; Accepted 22-04-2025; Published 29-04-2024

# Evaluation of Prompting Strategies for Cyberbullying Detection Using Various Large Language Models

**Anamika Gupta** anamikargupta@sscbsdu.ac.in
*Shaheed Sukhdev College of Business Studies*
*University of Delhi*
*New Delhi, India*

**Sakshi Garg** sakshi.23726@sscbs.du.ac.in
*Shaheed Sukhdev College of Business Studies*
*University of Delhi*
*New Delhi, India*

**Harsh Bamotra** harshbamotra.andc.du@gmail.com
*Acharya Narendra Dev College*
*University of Delhi*
*New Delhi, India*

**Corresponding Author:** Anamika Gupta



## Abstract

Sentiment analysis detects toxic language for safer online spaces and helps businesses refine strategies through customer feedback analysis [1, 2]. Advancements in Large Language Models (LLMs) and prompt engineering have introduced novel approaches to sentiment analysis, cyberbullying detection, and toxicity classification. However, several challenges persist, particularly in handling text ambiguity, sarcasm, multilingual contexts, and nuanced emotional comprehension, which limit the ability to achieve accurate and human-aligned results. This study uses the CYBY23 dataset, which contains 112 human-annotated threads. To balance the dataset, synthetic threads were generated using ChatGPT, resulting in a final dataset of 148 threads evenly distributed across two labels: 0 (bullying with no aggression) and 1 (bullying with aggression). Three publicly available LLMs—Deepseek-r1-distill-llama-70b (Deepseek), Qwen-2.5-32b (Qwen) and llama3-70b-8192 (Llama)—were systematically evaluated using zero-shot, one-shot, and few-shot prompting strategies, with all models accessed via Groq Cloud APIs. The model outputs were assessed using recall, precision, F1 scores, and accuracy to measure performance in different prompting techniques (PT). In this report, Qwen achieved the highest overall accuracy at 82.43% in few-shot 2, while Llama matched that accuracy in one-shot 2, demonstrating solid performance in few-shot tasks as well. Deepseek showed high variability, thriving with contextual enhancements in zero-shot 2 but struggling in one-shot and fluctuating in few-shot settings. one-shot prompting proved most effective for Llama, while few-shot methods worked best for both Qwen and Llama.

**Keywords:** Large-Language Models, Cyberbullying, Prompting Techniques, Aggression Detection, Sentiment Analysis



**Citation:** Gupta et al. Evaluation of Prompting Strategies for Cyberbullying Detection Using Various Large Language Models. Advances in Knowledge-Based Systems, Data Science, and Cybersecurity. Research 2025;2(1):9. Pages-184-196

## 1. INTRODUCTION

The term ”cyberbullying” was first introduced by Bill Belsey, a Canadian educator and anti-bullying advocate [3, 4]. Since its inception, cyberbullying has been explored extensively, with various interpretations provided by researchers [5–8]. At its core, cyberbullying is characterized as aggressive, abusive behavior that is intentionally and repeatedly directed toward individuals or groups through electronic media. This often involves the dissemination of offensive content or participation in other forms of social violence. A critical component of understanding and addressing cyberbullying is sentiment analysis, which involves the identification and categorization of emotions—positive, negative, or neutral. Beyond its general applications, sentiment analysis extends to the detection of toxic language, including hate speech, abuse, and racism [9].

Sentiment analysis is not limited to addressing cyberbullying but serves a multifaceted role in other domains. For example, during times of war or conflict, sentiment analysis offers valuable insight into opposing viewpoints, facilitating peaceful resolutions [10]. On social media platforms, it helps to gauge public opinion, enabling companies and content creators to refine their strategies. In addition, sentiment analysis empowers individuals to identify areas of concern, such as responding to negative comments and provides actionable insights for customer engagement. In marketing, sentiment analysis has proven instrumental in evaluating electronic word-of-mouth (eWOM), which refers to consumer-generated reviews and opinions shared online. eWOM significantly influences purchasing decisions, aiding financial institutions and companies in tailoring their strategies to better align with consumer needs and preferences [11]. With the rapid growth of social media, the dynamics of cyberbullying have evolved. Features such as accessibility (e.g. the ability to locate a target), information retrieval (e.g., gathering data about a target), editability (e.g., editing or deleting posts to deny involvement), and association (e.g., attributing blame to others) have contributed to the increasing prevalence of cyberbullying on these platforms [12].To address these challenges, numerous approaches have been used to analyze sentiments and detect toxicity in text. These include machine learning [13–15], natural language processing [16], deep learning [17], lexicon-based methods [18], and more recently, large language models (LLMs) [19–23]. Among these techniques, deep learning has shown promise in automating feature extraction from text, eliminating the need for manual feature engineering. It can leverage vast amounts of unlabeled data, enhancing the model’s generalization capabilities. However, deep learning models often require substantial labeled data for training, and the process can be complex and time-consuming [22]. Further, GPT-based models have emerged as a breakthrough in sentiment analysis. For instance, SentimentGPT has demonstrated significant improvements over traditional machine learning solutions for analyzing social media posts, particularly tweets [24, 25].

The advent of state-of-the-art large language models (LLMs), such as ChatGPT by OpenAI [26], Gemini by Google [27], Claude by Anthropic [28], and Llama by Meta [29], has revolutionized diverse domains. These advanced generative AI systems enable tasks such as data analysis [30], content creation [31], code translation [32], and fraud detection [33]. Despite their transformative capabilities, sentiment and toxicity analysis present persistent challenges. Issues such as ambiguity and contextual dependence, where words carry multiple meanings, and sarcasm and irony, where literal and intended meanings diverge, complicate the analysis. Moreover, multilingual and bilingual texts introduce further complexity, as variations in grammar, idioms, and cultural contexts significantly influence sentiment interpretation [34]. Additionally, as noted by Jamin Rahman Jim et al. [35], the use of emojis and emoticons adds another layer of difficulty in interpreting text-

based emotions. These visual symbols can convey sentiments that may not align with the accompanying text, thereby complicating accurate sentiment analysis.The inherent limitations of LLMs in comprehending nuanced human emotions and subtle linguistic cues pose significant challenges, particularly for tasks like toxicity detection that require precise contextual understanding [19]. Recent advancements in prompt engineering techniques, including zero-shot, one-shot, few-shot, and chain-of-thought prompting [36], have shown promise in optimizing LLM performance through structured and contextually rich instructions. However, the extent to which LLMs & prompting techniques achieve human-aligned toxicity analysis remains uncertain.

To address these gaps, this study undertakes a comparative evaluation of the performance of multiple state-of-the-art LLMs, including Deepseek-r1-distill-llama-70b, qwen-2.5-32b ,and llama3-70b-8192 using diverse prompting techniques such as zero-shot, one-shot, and few-shot approaches on the CYBY23 dataset [37].The CYBY23 dataset comprises 639 tweets organized into 112 threads, where each thread contains a main tweet followed by its replies. These threads have been meticulously annotated by human raters to determine aggression levels, offering a robust benchmark for assessing LLM capabilities. By comparing LLM-generated outputs with human judgments, this report addresses critical limitations in prompt engineering for nuanced language tasks. The findings contribute valuable perspectives on enhancing the real-world effectiveness of LLMs in sentiment and toxicity analysis, advancing their refinement and utility for addressing sensitive and challenging scenarios.

### 1.1 Objectives

- Utilized human-annotated datasets to compare LLM-generated outputs and analyze the effectiveness of LLMs in comprehending and classifying bullying and aggression in tweets or texts.
- Performance analysis of LLMs— Deepseek, Qwen ,and Llama—across a few prompting techniques: zero-shot, one-shot, and few-shot approaches.
- Examined the impact of changing examples in prompting techniques on the accuracy of LLMs.
- Assessment of how rules, definitions, and evaluation criteria in prompting techniques influence LLM responses in interpreting toxicity.

## 2. METHODOLOGY

In this study, we conducted a comparative analysis of how different Large Language Models (LLMs) interpret toxicity and toxic behavior in a human-annotated dataset by utilizing the CYBY23 dataset. The dataset comprises 639 tweets organized into 112 threads. Each thread consists of a main tweet and its corresponding replies across three distinct aggression labels: 0 (No bullying, only aggression), 1 (Bullying with low aggression), and 2 (Bullying with high aggression). These labels were assigned through a majority vote process involving annotators from diverse cultural backgrounds, ensuring the objectivity and reliability of the labeling. Annotators were recruited from different age groups and cultural backgrounds, with English proficiency as a mandatory requirement to match the language of the data. Each thread was independently labeled by five annotators,

and only threads where a consensus among annotators was achieved were retained in the final dataset [37]. The dataset contains 38 instances of label 0, 61 instances of label 1, and 13 instances of label 2. To simplify the distinction between no bullying, only aggression, and bullying with aggression, all label-2 instances are merged into label 1, increasing its count to 74. As shown in FIGURE 1, OpenAI's ChatGPT was used to generate synthetic threads for label 0, to balance the number of instances between classes 0 and 1. A similar approach of generating synthetic data using GPT models was explored by Busker et al. [38], where they studied how varying the granularity of prompts influenced the characteristics of the generated data. After augmentation, the dataset expanded to 148 threads, evenly split: 74 instances of label 0 (bullying without aggression) and 74 of label 1 (bullying with aggression).

The dataset is then processed through three different phases as presented in FIGURE 2. The first phase focuses on data preprocessing and model selection, where the threads (comprising the main tweet and its corresponding replies) are formatted and passed as input to the selected LLMs. In the second phase, prompting strategies are developed, with each model evaluated using three different techniques—zero-shot, one-shot, and few-shot—across three LLMs: Deepseek-r1-distill-llama-70b by Deepseek, qwen-2.5-32b by Alibaba Cloud, and llama3-70b-8192 by Meta. The third phase, the results phase, involves analyzing and comparing the outputs generated by each model under different prompting strategies against the ground-truth labels in the dataset to assess performance metrics.

| Scale | Class Label | Percentages | Value Count |
|---|---|---|---|
| 0 | No bullying, only aggression | 50% | 74 |
| 1 | Bullying with aggression | 50% | 74 |

Figure 1: Thread counts corresponding to three aggression levels in the main dataset -0(No bullying only aggression), 1(bullying with low aggression) and 2(Bullying with high aggression)

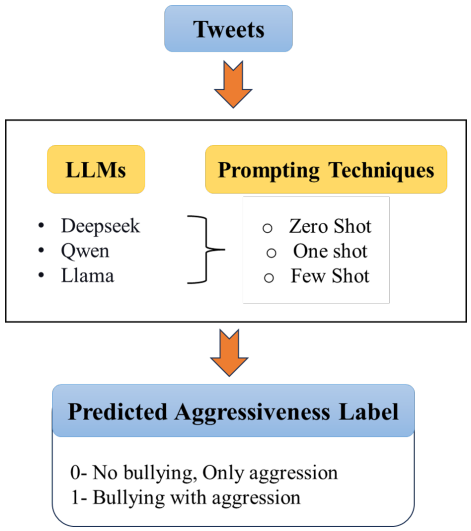


Figure 2: Process flow for the comparative analysis of performance metrics of three LLMs—Deepseek, Qwen, and Llama—in predicting aggression levels:0 (No bullying, only aggression) 1 (Bullying with aggression)

### 2.1 Data Pre-processing and Model Selection

In data preprocessing, the models were instructed to treat the first line of each entry as the main tweet, with subsequent lines starting with '@' considered reply tweets. However, since some replies in the original dataset did not begin with '@', it was manually added to the start of all reply tweets for consistency. Furthermore, when the main tweet began with an asterisk (*), it was removed to maintain consistent formatting, as illustrated in FIGURE 3. These preprocessing steps were designed to help the model better interpret the data structure and respond more accurately to the prompts. The preprocessed data was then passed to the models, with priority given to those that were publicly available, hence the models Llama, Gwen, and Deepseek are chosen and accessed via the Groq Cloud API and evaluated using the defined prompting techniques.

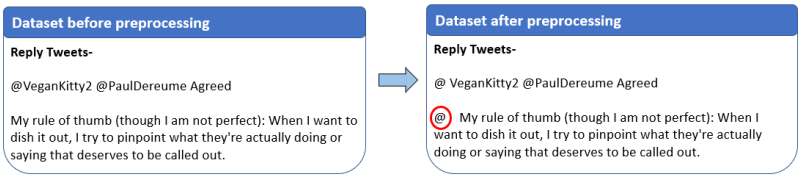


Figure 3: Adding '@' in front of all reply tweet to maintain consistency

### 2.2 Prompting Strategy

To comparatively analyze the responses of different models to aggression, three prompting strategies were used: zero-shot [39], one-shot [36], and few-shot [40], to assess their effectiveness in improving the model's accuracy in predicting and interpreting aggression.

In zero-shot prompting technique, two distinct prompts were created. In the first prompt(zero-shot 1), illustrated in FIGURE 4(a), we provided the two labels—0: No bullying, only aggression, 1: Bullying with aggression. The second prompt(zero-shot 2) presented a deeper analysis of the dataset to identify patterns associated with each label, which we incorporated into the prompt as "Characteristics." These patterns were supplemented with four specific instructions, referred to as "critical evaluation criteria," designed to address gaps in the model's ability to accurately interpret and differentiate nuanced levels of aggression and bullying, as shown in FIGURE 4(b). Subsequently, the one-shot prompting technique, in contrast to zero-shot prompting, incorporated a single demonstration alongside the task-specific instruction prompt. As illustrated in FIGURE 5, the base prompt remained unchanged, while different example inputs were introduced for labels 0 and 1. In One-Shot 1, the prompt from FIGURE 5(a) was paired with the example in FIGURE 5(b), whereas in One-Shot 2, Figure 5(a) was combined with the example in FIGURE 5(c). Whereas zero-shot prompting emphasized the identification of distinguishing characteristics for each label, one-shot prompting prioritized the recognition of patterns indicative of bullying and aggression. After analyzing the parameters that the model struggled to comprehend, we introduced specific rules as mentioned in FIGURE 5(a), to enhance its understanding. To improve clarity and task alignment, we structured the rules to be precise and descriptive while restricting the expected outputs strictly to the predefined labels (0 and 1).

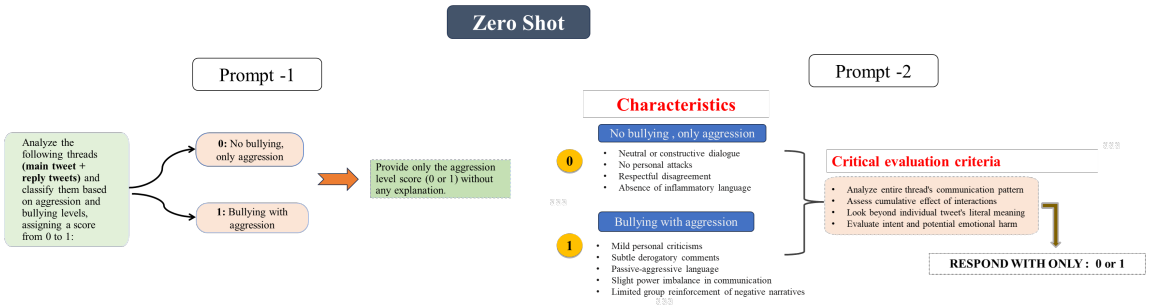


Figure 4: Two distinct zero-shot prompting techniques associated with two classification 0: No bullying, only aggression Aggression without signs of bullying, 1: Bullying with aggression

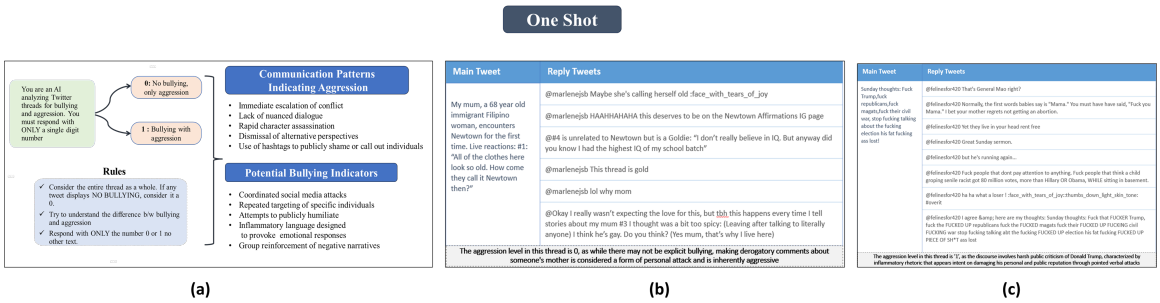


Figure 5: One shot prompt supplemented with example of label 0 and label 1

In the case of few-shot prompting, five different prompts were examined to evaluate their effectiveness in enhancing model performance. As demonstrated in FIGURE 6(a), the first prompt(few-shot 1) incorporated definitions of bullying from Alfurayj et al. [37] and definitions of aggression from Google Dictionary. These definitions were paired with examples corresponding to each label (0 and 1), as illustrated in FIGURES 5(b) and 5(c). The second prompt(few-shot 2) is identical to FIGURE 5(a), and was supplemented with examples for each label, as shown in FIGURES 6(c) and 6(d). The third prompt(few-shot 3) adhered to the same structural framework outlined in FIGURE 5. All these examples were carefully selected from the dataset, prioritizing those with multiple reply tweets to assess how effectively the model interprets toxicity when provided with the maximum available content(tweets). The fourth prompt(few-shot 4) combined the zero-shot prompt (FIGURE 4(b)) with examples from FIGURES 6(c) and 6(d). Finally, the fifth prompt(few-shot 5), as referenced in FIGURE 6(b), was supplemented with the examples used in FIGURES 5(b) and 5(c).

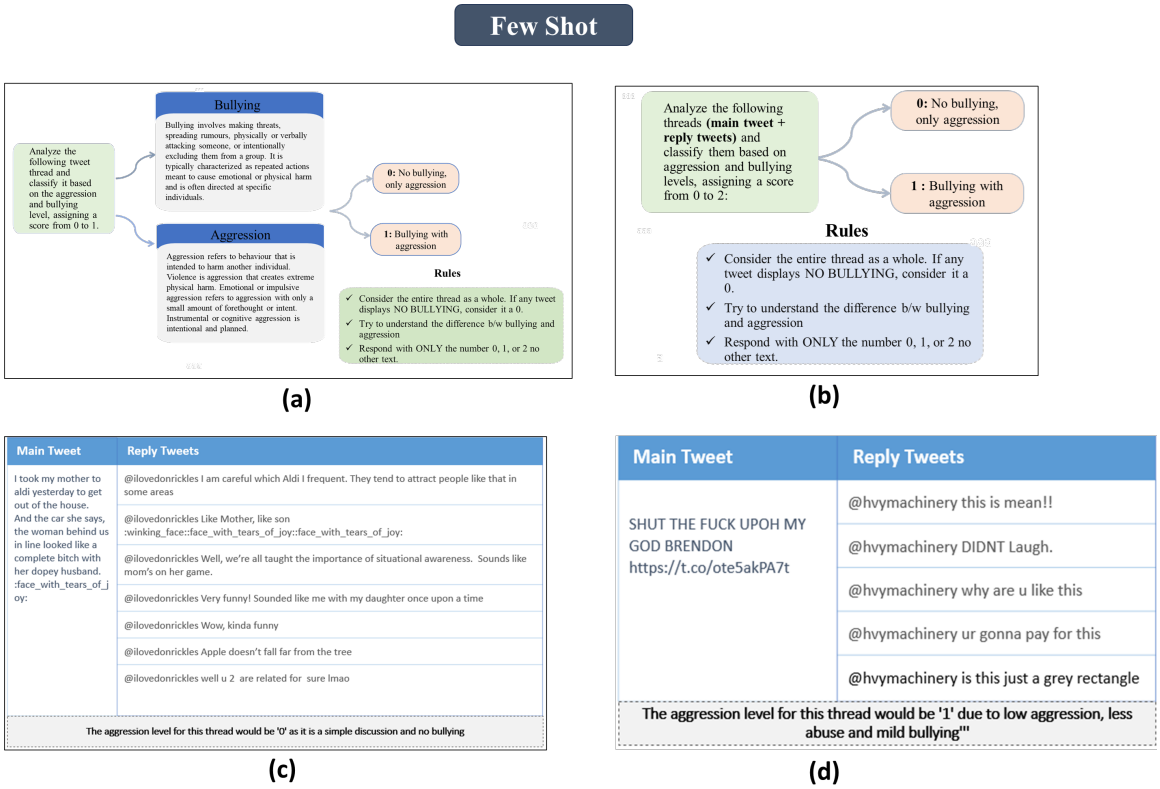


Figure 6: Few shot prompts with examples corresponding to three distinct labels—0: No bullying, only aggression Aggression without signs of bullying, 1: Bullying with aggression

## 3. RESULTS AND DISCUSSION

The performance of the models Deepseek, Llama and Qwen was evaluated using zero-shot, one-shot, and few-shot prompting techniques, with performance metrics summarized in TABLE 1. Four evaluation metrics were employed to assess the performance of the LLMs: Precision, recall, F1 score and accuracy [41]. The accuracy score of the prompting techniques in different models is visualized through a heat map, as illustrated in FIGURE 7.

| PT | Deepseek | | | Qwen | | | Llama | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| zero-shot 1 | 0.6696 | 0.6486 | 0.6375 | 0.8008 | 0.7973 | 0.7967 | 0.6339 | 0.6284 | 0.6245 |
| zero-shot 2 | 0.7975 | 0.7635 | 0.7566 | 0.7718 | 0.6757 | 0.6442 | 0.6701 | 0.6622 | 0.6582 |
| one-shot 1 | 0.6794 | 0.6216 | 0.5885 | 0.7914 | 0.7838 | 0.7824 | 0.8008 | 0.7973 | 0.7967 |
| one-shot 2 | 0.6897 | 0.5811 | 0.5111 | 0.7478 | 0.7297 | 0.7247 | 0.8304 | 0.8243 | 0.8235 |
| few-shot 1 | 0.7753 | 0.7703 | 0.7692 | 0.8171 | 0.8041 | 0.8020 | 0.7721 | 0.7703 | 0.7699 |
| few-shot 2 | 0.7075 | 0.6892 | 0.6822 | 0.8253 | 0.8243 | 0.8242 | 0.7575 | 0.7568 | 0.7566 |
| few-shot 3 | 0.7735 | 0.6486 | 0.6034 | 0.7639 | 0.7500 | 0.7467 | 0.7812 | 0.7770 | 0.7762 |
| few-shot 4 | 0.7771 | 0.7770 | 0.7770 | 0.8028 | 0.7973 | 0.7964 | 0.7835 | 0.7703 | 0.7676 |
| few-shot 5 | 0.7643 | 0.6622 | 0.6260 | 0.7910 | 0.7905 | 0.7905 | 0.7647 | 0.7635 | 0.7632 |

Table 1: Performance Metrics for Deepseek, Qwen, and Llama corresponding to three prompting techniques (PT): Zero Shot, One Shot, and Few Shot

| PT | Deepseek | Qwen | Llama |
|---|---|---|---|
| zero-shot 1 | 0.6486 | 0.7973 | 0.6284 |
| zero-shot 2 | 0.7635 | 0.6757 | 0.6622 |
| one-shot 1 | 0.6216 | 0.7838 | 0.7973 |
| one-shot 2 | 0.5811 | 0.7297 | 0.8243 |
| few-shot 1 | 0.7703 | 0.8041 | 0.7703 |
| few-shot 2 | 0.6892 | 0.8243 | 0.7568 |
| few-shot 3 | 0.6486 | 0.7500 | 0.7770 |
| few-shot 4 | 0.7770 | 0.7973 | 0.7703 |
| few-shot 5 | 0.6622 | 0.7905 | 0.7635 |

Table 2: Accuracy Comparison for Deepseek, Qwen, and Llama across Prompting Techniques

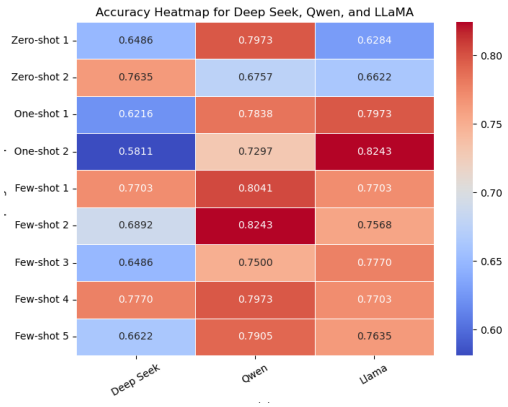


Figure 7: Heat map of F1 score of prompting techniques across three models: Deepseek, Qwen, and Llama

The results in TABLE 1 and TABLE 2, highlight the variability in performance when using the mentioned prompting techniques. In case of zero-shot prompting, performance metrics indicate that Qwen processes direct inputs more effectively than the other models, utilizing its pre-trained knowledge to infer classifications, obtaining an accuracy and recall of 79.73% , F1 score of 79.67% and precision of 80.08% . However, Llama struggled with accuracy, obtaining the lowest accuracy of 62.84% ,F1 score of 62.45% and precision of 63.39% in the zero-shot 1 scenario. Interestingly, performance varied significantly in zero-shot 2, where contextual cues and characteristics associated with labels were incorporated. Deepseek and Llama benefited from contextual enrichment, leading to an F1 score improvement from 63.75% to 75.66% for Deepseek and from 62.45% to 65.82% for Llama. These findings indicate that while zero-shot prompting can be effective, contextual enhancements are particularly beneficial for Deepseek and Llama, whereas Qwen may leverage direct label cues more effectively.

The one-shot prompting experiments highlight differences in how each model responds to a single example when making predictions. LLaMA shows the most significant improvement, with accuracy and F1 scores rising from 79.73% to 82.43% in One-Shot 1 and from 79.67% to 82.35% in One-Shot 2 when provided with a label 1 example. This suggests that the model may struggle to interpret aggression-driven bullying without explicit guidance, relying heavily on contextual cues from the given example to refine its classification. In contrast, the performance of the Qwen drop in F1 from 78.24% in one-shot 1 to 72.47% in one-shot 2 suggests that the introduction of an example of label 1 was not particularly effective for this model. Deepseek, meanwhile, struggles with One-Shot prompting, as its accuracy declines from 64.86% to 58.11%, along with a decrease in F1 score when given a label 1 example. This implied that the model may not be effectively leveraging the provided example and could be misclassifying more instances as a result. Few-shot prompting results denote that increasing the number of examples generally improves model performance, but the extent of improvement varies across models. Llama achieves consistent performance across multiple few-shot scenarios, with values reaching 77.70% in few-shot 1 (77.03%) and few-shot 3 (77.70%), showing that it effectively leverages multiple examples to refine its predictions. Qwen also benefits from few-shot prompting, achieving an accuracy of 82.43% in few-shot 2, highlighting that it generalizes well when given more context. Deepseek demonstrates fluctuating recall and F1 scores, suggesting that it does not always benefit from additional examples in a structured manner. Notably, Llama's accuracy slightly decreases compared to its one-shot performance in certain cases, indicating that while a single well-chosen example works well, excessive contextual information might not always be beneficial. On the other hand, Qwen maintains a stable recall and F1 score across different few-shot settings, reinforcing its ability to understand multiple examples.

From a comparative perspective on prompting strategies, we observed distinct adaptation patterns among Deepseek, Qwen, and Llama. Qwen consistently excels, achieving the highest F1 score (82.42%) in few-shot 2 and effectively integrating multiple examples. Llama demonstrates strong one-shot learning, reaching 82.35% in one-shot 2, and maintains high performance in few-shot settings. Deepseek shows higher variability, struggling in one-shot but benefiting from contextual enhancements in zero-shot 2, with its best performance (77.70%) in few-shot 4. Overall, Llama performs best with one-shot prompting, while few-shot yields optimal results for both Qwen and Llama. Deepseek may benefit from a more refined example selection to enhance its performance.

Differences in the performance behavior of models across all performance parameters—accuracy, precision, recall, and F1 score—for different prompt techniques (few-shot, one-shot, and multi-shot) are likely attributable to differences in pretraining data, model architecture, and hyperparameter tuning. Moreover, performance differences across prompting types underscore that strategy directly influences outcomes, with each model responding uniquely to the quantity and nature of the context provided. Considering models' sensitivity to prompt strategies, future research could focus on designing prompts that are easily adaptable and consistently yield high performance for multiclass datasets—particularly those aimed at better distinguishing between closely related labels: label 1 (bullying with low aggression) and label 2 (bullying with high aggression), as well as additional labels reflecting varying intensities of aggression. This would help to determine aggression with more precision.

## 4. CONCLUSION

In this study, we conducted a comparative analysis of how different Large Language Models (LLMs) interpret bullying and aggression using various prompting strategies. We evaluated Deepseek-r1-distill-llama-70b, qwen-2.5-32b , and llama3-70b-8192 across zero-shot, one-shot, and few-shot prompting techniques. zero-shot prompting included both a basic label-based approach and an enhanced version with contextual cues. one-shot prompting introduced structured rules with a single example, while few-shot prompting incorporated multiple examples and definitions to refine model predictions. Results show that Qwen performed best overall, achieving the highest accuracy (82.43%) in few-shot 2. Llama excelled in one-shot prompting, reaching 82.43% in one-shot 2, and maintained strong few-shot performance. Deepseek showed higher variability, benefiting from contextual enhancements in zero-shot 2 but struggling in one-shot setting. In few-shot prompting, its performance fluctuated, indicating that it did not consistently benefit from multiple examples and required optimized selection for improvement. Overall, one-shot prompting was most effective for Llama, while few-shot worked best for Qwen and Llama. Deepseek required more structured example selection for optimal results.

## References

[1] S. Feuerriegel et al., "Generative AI," *Business Information Systems Engineering*, vol. 66, no. 1, pp. 111–126, 2024.

[2] S. Ahmadi, "Open AI and its Impact on Fraud Detection in Financial Industry," *Journal of Knowledge Learning and Science Technology ISSN*, pp. 2959–6386, 2023.

[3] S. Bauman, "Cyberbullying: A virtual menace," in *National Coalition Against Bullying National Conference*, vol. 2, no. 4, 2007.

[4] M. Campbell and S. Bauman, "Cyberbullying: Definition, Consequences, Prevalence," in *Reducing Cyberbullying in Schools*, pp. 3–16, Elsevier, 2018.

[5] K. T. A. S. Kasturiratna et al., "Umbrella review of meta-analyses on the risk factors, protective factors, consequences and interventions of cyberbullying victimization," *Nature Human Behaviour*, pp. 1–32, 2024.

[6] B. S. Nandhini and J. I. Sheeba, "Online social network bullying detection using intelligence techniques," *Procedia Computer Science*, vol. 45, pp. 485–492, 2015.

[7] S. P. Kiriakidis and A. Kavoura, "Cyberbullying: A review of the literature on harassment through the internet and other electronic means," *Family Community Health*, vol. 33, no. 2, pp. 82–93, 2010.

[8] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Computer Science*, vol. 181, pp. 605–611, 2021.

[9] G. H. Resende et al., "A Comprehensive View of the Biases of Toxicity and Sentiment Analysis Methods Towards Utterances with African American English Expressions," *arXiv preprint arXiv:2401.12720*, 2024.

[10] A. Liyih et al., "Sentiment analysis of the Hamas-Israel war on YouTube comments using deep learning," *Scientific Reports*, vol. 14, 2024, Nature Publishing Group UK London.

[11] M. Mathebula, A. Modupe, and V. Marivate, "ChatGPT as a Text Annotation Tool to Evaluate Sentiment Analysis on South African Financial Institutions," *IEEE Access*, vol. 12, pp. 144017–144043, 2024.

[12] G. Giumetti and R. Kowalski, "Cyberbullying via Social Media and Well-Being," *Current Opinion in Psychology*, vol. 45, p. 101314, 2022.

[13] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1–5, IEEE, 2013.

[14] Gupta A., Thakkar K., Bhasin V., Mathur V., and Tiwari A. (2024), "Bystander Detection: Automatic Labeling Techniques using Feature Selection and Machine Learning" *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 15, issue 1, pp. 1135–1143, 2024. http://dx.doi.org/10.14569/IJACSA.2024.01501112

[15] Gupta, A., Thakkar, K., Mathur, V., and Tiwari, A. (2023). Machine Learning Methods for Detection of Bystanders: A Survey. *International Journal of Advanced Computer Technology*, vol. 12, issue 4, pp. 06–14, 2024.

[16] M. R. Hasan, M. Maliha, and M. Arifuzzaman, "Sentiment analysis with NLP on Twitter data," in *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, pp. 1–4, IEEE, 2019.

[17] Q. T. Ain et al., "Sentiment analysis using deep learning techniques: a review," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017.

[18] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 81, 2021.

[19] W. Zhang et al., "Sentiment Analysis in the Era of Large Language Models: A Reality Check," 2023. [Online]. Available: `https://doi.org/10.48550/arXiv.2305.15005`.

[20] M. S. Islam, S. Sutton, and R. I. Rafiq, "A Generative AI Powered Approach to Cyberbullying Detection," in *Proceedings of the 2024 8th International Conference on Information System and Data Mining*, pp. 57–63, 2024.

[21] K. Verma et al., "Beyond Binary: Towards Embracing Complexities in Cyberbullying Detection and Intervention - a Position Paper," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2264–2284, 2024.

[22] T. Zhan et al., "Optimization techniques for sentiment analysis based on LLM (GPT-3)," *Applied and Computational Engineering*, vol. 67, pp. 41–47, 2024.

[23] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification," *Multimedia Systems*, vol. 28, no. 6, pp. 1897–1904, 2022.

[24] K. Kheiri and H. Karimi, "SentimentGPT: Exploiting GPT for advanced sentiment analysis and its departure from current machine learning," *arXiv preprint arXiv:2307.10234*, 2023.

[25] G. Villate-Castillo, J. D. Ser, and B. S. Urquijo, "A systematic review of toxicity in large language models: Definitions, datasets, detectors, detoxification methods and challenges," 2024.

[26] OpenAI et al., "GPT-4 Technical Report," 2023. [Online]. Available: https://arxiv.org/abs/2303.08774.

[27] S. Pichai, "Our next-generation model: Gemini 1.5," 2024. [Online]. Available: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note.

[28] "The Claude 3 Model Family: Opus, Sonnet, Haiku," 2024. [Online]. Available: https://paperswithcode.com/paper/the-claude-3-model-family-opus-sonnet-haiku.

[29] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[30] J. P. Inala et al., "Data Analysis in the Era of Generative AI," *arXiv preprint arXiv:2409.18475*, 2024.

[31] R. Wahid, J. Mero, and P. Ritala, "Written by ChatGPT, illustrated by Midjourney: generative AI for content marketing," *Asia Pacific Journal of Marketing and Logistics*, vol. 35, no. 8, pp. 1813–1822, 2023.

[32] J. D. Weisz et al., "Perfection not required? Human-AI partnerships in code translation," in *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 402–412, 2021.

[33] D. Krause, "Mitigating risks for financial firms using generative AI tools," *Available at SSRN 4452600*, 2023.

[34] S. Gupta, R. Ranjan, and S. N. Singh, "Comprehensive Study on Sentiment Analysis: From Rule-based to modern LLM based system," *arXiv preprint arXiv:2409.09989*, 2024.

[35] J. R. Jim et al., "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, p. 100059, 2024, Elsevier.

[36] B. Chen et al., "Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review," *arXiv preprint arXiv:2310.14735*, 2023.

[37] H. S. Alfurayj, N. S. Yee, and S. L. Lutfi, "Bystanders Unveiled: Introducing a Comprehensive Cyberbullying Corpus with Bystander Information," in *TENCON 2023-2023 IEEE Region 10 Conference (TENCON)*, pp. 1012–1017, IEEE, 2023.

[38] Busker, T., Choenni, S., Bargh, M. S. (2025). Exploiting GPT for synthetic data generation: An empirical study. Government Information Quarterly,42(1), 101988

[39] P. Sahoo et al., "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," 2024. [Online]. Available: https://doi.org/10.13140/RG.2.2.13032.65286.

[40] T. Brown et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[41] M. Hossin and M. N. Sulaiman, “A Review on Evaluation Metrics for Data Classification Evaluations,” in *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 1, Academy & Industry Research Collaboration Center (AIRCC), 2015.