2024 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-7488-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMLA61862.2024.00257

Detecting Cyberbullying in Visual Content: A Large Vision-Language Model Approach

Jaden Mu[‡][§], David Cong[¶][§], Helen Qin[‖][§], Ishan Ajay[††][§], Keyan Guo[*], Nishant Vishwamitra[†], Hongxin Hu[*]
[*]University at Buffalo, [†]University of Texas at San Antonio, [‡]East Chapel Hill High School,
[¶]Williamsville East High School, [‖]Thomas Jefferson High School for Science and Technology, [††]John Glenn School

***Abstract*—Cyberbullying has rapidly evolved with the evolution of online platforms, transcending traditional text-based forms to include images and other multimedia content. Two major challenges are identified in detecting cyberbullying images: recognizing cyberbullying-related visual factors and addressing the context-dependent nature of such images. In this paper, we conduct a comprehensive investigation of the ability of Large Vision-Language Models (LVLMs) to evaluate visual factors related to cyberbullying, and to interpret the context-dependent nature of such images. Furthermore, by proposing a diverse set of prompting strategies, we optimize LVLMs for cyberbullying image detection. In particular, through our carefully crafted Chain-of-Thought (CoT) methodology, we guide the model through structured reasoning pathways to interpret complex visual factors and account for their context. Our results show that the structured reasoning pathways significantly enhance model performance, achieving state-of-the-art accuracy and precision while remaining efficient by eliminating the need for any extensive training process.**

Disclaimer. This paper contains harmful content, which may offend or disturb readers.

I. Introduction

Cyberbullying, which is the use of digital technologies to harass, threaten, or embarrass individuals, has become a widespread concern with the rapid proliferation of social media and online communication platforms. Unlike traditional bullying, cyberbullying can occur anytime, and with perpetrators often being anonymous, is particularly difficult for victims to escape. Studies show that cyberbullying affects a significant portion of young people globally, with surveys indicating that 37% of teens in the United States have experienced online harassment, and around 15% have been the target of severe forms of cyberbullying, such as threats of violence [1]. The psychological impacts of cyberbullying on young people can be devastating, leading to anxiety, depression, low self-esteem, and, in extreme cases, self-harm or suicide [2], [3].

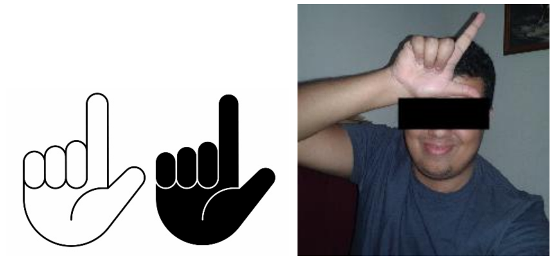
Cyberbullying has traditionally been associated with text-based content. However, the rapid rise of image-based social media platforms such as Instagram [4], Snapchat [5], and TikTok [6] has expanded the scope of online harassment. Images and multimedia content often carry implicit messages that can be just as harmful, if not more so, than written words. For instance, individuals can be targeted through altered photos, humiliating memes, or gestures that convey derogatory or threatening intent. Studies indicate that visual content can amplify the emotional impact of cyberbullying, making it more personal and damaging, particularly among young users [7], [8]. Furthermore, the detection of cyberbullying in images poses significant challenges for traditional machine learning models, as it requires understanding complex, context-dependent factors such as facial expressions, body language, and symbolic gestures [9]. As visual content continues to dominate social media interactions, addressing cyberbullying in images has become critical for developing effective safety solutions. While extensive research has focused on using machine learning and artificial intelligence (ML/AI) techniques to identify cyberbullying within images, certain gaps still remain text [10]–[14],

Cyberbullying image detection presents unique challenges that go beyond those faced in text-based detection. *First*, a major challenge lies in accurately identifying the visual factors that contribute to cyberbullying. Unlike text, where harmful content can often be explicitly detected through keywords or phrases, images require the recognition and interpretation of multiple visual elements such as facial expressions, body language, gestures, and objects. These elements can vary widely across different contexts, making the task of detecting bullying content in images more complex. For example, a smile may be benign in one context but mocking in another, depending on the surrounding visual and social cues. *Second*, the context-dependent nature of cyberbullying images makes detection more difficult. The interpretation of these visual factors often depends heavily on context, which can be influenced by cultural, societal, and relational factors. An visual factor might not be inherently related to bullying on its own, but when paired with specific context, can convey a powerful message of harassment or intimidation. This context-dependent nature means that detection models need to not only recognize individual visual elements, but also understand the relationships between these elements within a given context to accurately detect cyberbullying.

Recent advances in large vision-language models (LVLMs) offer a promising solution to the complex task of cyberbullying image detection. These models integrate both visual and textual data, enabling them to better understand the nuanced and context-dependent nature of visual content. LVLMs have demonstrated remarkable success in a variety of tasks that require reasoning across modalities [15]–[17], and their application to cyberbullying detection holds significant potential. Unlike traditional machine learning methods, which often require extensive labeled datasets and preprocessing,

[§] Work done during internship at University at Buffalo.

1946-0759/24/$31.00 ©2024 IEEE
DOI 10.1109/ICMLA61862.2024.00257



(a) Gesture (b) Gesture with Context

Fig. 1: Inter-factor Relationship in Cyberbullying Images

LVLMs can adapt to new tasks more efficiently through prompt-based adaptation, even with limited domain-specific data. However, their use in detecting cyberbullying in images remains underexplored, which is a gap this paper aims to bridge.

Recognizing the complexity of this problem, we leverage the inherent capabilities of LVLMs to process both visual and textual data in a unified manner by designing a diverse set of prompts that adapt the LVLM to cyberbullying image detection. We introduce task-specific prompts that guide the models to focus on critical visual factors such as facial expressions, gestures, body language, and social factors. Additionally, we structure reasoning pathways to encourage the model to break down complex relationships within the image and infer whether the visual content constitutes cyberbullying. Through a series of carefully crafted experiments, we demonstrate that this prompt-driven adaptation significantly enhances the model's performance. Our approach not only explores the capability of LVLMs in cyberbullying detection but also enhances the accuracy and precision of cyberbullying detection, surpassing previous state-of-the-art (SOTA) approaches. Furthmore, unlike traditional methods that rely on extensive labeled data and time-consuming training, our LVLM-based solution is able to adapt quickly through prompt engineering, offering a scalable and resource-efficient alternative while maintaining superior performance.

II. Preliminaries

A. The Nature of Cyberbullying Images

Cyberbullying images harass users on social media platforms by visually conveying threatening, embarrassing, or harmful messages [18]. Although these images may not contain explicit textual content, they communicate aggression through visual cues, making detection a challenging task. In this section, we introduce two major natures that are significant for detecting cyberbullying images.

Visual factors of cyberbullying images: Based on insights from a recent study [9], we identify five key visual factors that are commonly associated with cyberbullying in images: The **body pose** factor refers to the orientation and posture of individuals depicted in images. A frontal body pose facing the viewer, in particular, is highly associated with cyberbullying since it suggests the message of the image is targeted toward the viewer. The **facial expression** factor refers to the emotion of any human subjects in an image. In the context of cyberbullying, perpetrators may use facial expressions to mock, intimidate, or belittle their victims. While joyful expressions might typically be associated with positive interactions, in cyberbullying, such expressions can take on a mocking or derisive tone, exacerbating the harmful impact of the image. The **object** factor focuses on the presence of threatening or harmful items within images. Objects such as guns, knives, or other weapons are commonly used in cyberbullying. However, the object factor is particularly diverse, as there are a variety of objects that can be interpreted as cyberbullying depending on the context of the image. The **gesture** factor focuses on hand gestures in the image. Negative or offensive gestures are particularly prevalent in cyberbullying images, such as the middle finger, thumbs down, or the "loser" sign. Finally, **social** factors encompass the broader societal and cultural symbols that may be present in cyberbullying images, such as hate speech or symbols that target specific groups.

The Context-dependent nature of cyberbullying factors: Cyberbullying in images is highly context-dependent and often relies on the interplay between the five previously identified visual factors. The meaning of an image can shift dramatically depending on the combination of elements such as body pose, facial expressions, and surrounding symbols. This complexity makes it challenging to detect cyberbullying purely by isolating individual visual cues. For example, in image 1a, a potentially demeaning gesture is presented abstractly and not directed toward the viewer, leading it not to be classified as cyberbullying. However, in image 1b, the same gesture is made by a person directly facing the camera, signaling that the gesture is likely intended to harass or belittle the viewer. This example highlights how the body pose factor provides context to the gesture factor, demonstrating the complex, interrelated nature of visual elements in cyberbullying images.

B. Data Collection and Annotation

In our study, we utilize the well-known benchmark cyberbullying image dataset developed by Vishwamitra et al. [9]. This large-scale, real-world dataset was compiled by scraping web search engines (Google [19], Bing [20], Baidu [21]) and social media platforms (Instagram [4], Flickr [22], X [23]) for images tagged with keywords related to cyberbullying. To ensure a representative and diverse sample, the dataset includes images from a wide range of browsers and platforms, capturing different contexts and scenarios where cyberbullying may occur. The final dataset comprises 19,300 images, each of which has been annotated as either "cyberbullying" or "non-cyberbullying" through a majority voting process involving three human annotators.

To assess whether LVLMs can effectively detect cyberbullying images, we randomly selected 1,000 cyberbullying images and 1,000 non-cyberbullying images as our test set. To further enhance this test dataset, three of our authors carefully evaluated each image based on the five visual factors identified earlier. Each annotator was given a set of specific questions

TABLE I: GPT-4 Accuracy in Evaluating Factors

Body Pose	Facial Expression	Gesture	Weapons	Social Issue
0.92	0.90	0.67	0.95	1.00

aimed at evaluating these factors, ensuring that all key elements of the image were considered. The questions presented to the annotators were designed to assess the following:

- Is there a person facing the viewer frontally? Yes/no.
- If there is a person, what is their facial expression (joy, sadness, disgust, contempt, anger, surprise, neutral)? If there is a facial expression not in the list, please specify.
- If there is a person making a gesture, what is the gesture (finger gun, loser symbol, middle finger, thumbs down)? If there is a gesture not in the list, please specify.
- If there is a threatening object in the image, what is it (gun, knife, noose)? If there is a threatening object not in the list, please specify.
- If the meaning of the image relevant to/affected by social context (e.g. racism, anti-LGBTQ), what is the specific context?

To standardize the labels, we further provide a predefined list of possible options for each question. However, annotators were also instructed to specify any expression, gesture, or object that was not included in the list, ensuring that the range of responses was not artificially constrained. In cases where there was disagreement between annotators, we resolved the issue by selecting the majority vote.

C. LVLMs in Identifying Visual Factors

With the annotated test dataset, we first aim to investigate whether LVLMs are capable of understanding and evaluating the five visual factors that dominate cyberbullying images. To this end, we conducted an experiment to measure the accuracy of the LVLM in evaluating these factors. In this experiment, the model was presented with 10 example images, including 5 cyberbullying and 5 non-cyberbullying, and tasked with evaluating each of the five visual factors for a new given image.

In the experiment, we choose GPT-4 [24], a well-know SOTA LVLM, as our evaluation model. As depicted in Table I, GPT-4 demonstrates strong performance in evaluating several key visual factors, achieving high accuracy in body pose at 0.92, facial expression at 0.90, the presence of dangerous objects at 0.95, and relevant social issues at 1.00. In particular, GPT-4 achieved perfect agreement with human annotators in identifying social issues, indicating a robust understanding of the contextual importance in cyberbullying images. However, GPT-4 exhibits lower accuracy of 0.67 when evaluating gestures, suggesting that the model struggles with interpreting more complex hand behaviors that may not be as straightforward as other visual cues. This limitation highlights the need for additional refinement in gesture recognition, particularly when gestures are context-dependent or vary across cultural and social contexts.

III. Prompting Strategies for Cyberbullying Image Detection

In this section, we examine the prompting strategies employed to adapt LVLMs to the task of cyberbullying image detection. These prompts were designed to guide the model in interpreting images by focusing on the visual factors identified earlier, allowing us to assess the LVLMs' capabilities in detecting cyberbullying content. Specifically, we evaluated three primary prompting strategies: the general prompt, few-shot prompt, and chain-of-thought prompt.

A. General Prompt

We first investigate if the LVLMs can identify given cyberbullying images in a zero-shot setting. We design a simple, general prompt that leverages a LVLM's instruction-following capabilities to adapt the model to the task of cyberbullying image detection. Additionally, we augment the general prompt with a formal definition of cyberbullying from [25] to provide it with necessary task-specific information.

General Prompt (GP): *Does the following image contain cyberbullying content? a. Yes b. No*

General Prompt with Definition (GPwDef): *Cyberbullying content is used to harass, threaten, embarrass, or target another person. Does the following image contain cyberbullying content? a. Yes b. No*

B. Few-Shot Prompt

While LLMs have demonstrated remarkable zero-shot capabilities [26], [27], their capabilities from general pretraining still fall short on more complicated tasks. Few-shot prompting is a validated prompting methodology to enable in-context learning [28], where labeled examples are provided to condition the model for future examples on a specific task [29], [30]. Building on previous work [30], we formulated our few-shot prompt by augmenting the general prompt with 10 randomly selected images — 5 classified as positive (cyberbullying) and 5 as negative (non-cyberbullying)—each paired with the corresponding label.

C. Chain-of-Thought Prompt

The Chain-of-Thought (CoT) prompting strategy [31], [32] enables complex reasoning by providing a series of intermediate steps and has been proven to achieve remarkable effectiveness in natural language processes (NLP) and computer vision (CV) downstream tasks [15]–[17]. These intermediate steps can also be specifically designed to provide the LLM with task-specific information. Because cyberbullying is strongly characterized by the five factors identified previously, we hypothesize that providing a structured reasoning pathway where the model must first evaluate the five factors and how they influence the final classification could improve performance. Therefore, we construct our CoT prompt by augmenting the general prompt (which is used to instruct the model to make the final classification) with a series of steps instructing the model to evaluate each of the five factors.

TABLE II: Accuracy in Detecting Listed Gestures

Without Verification Component	With Verification Component
0.79	0.88

Additionally, we add 10 examples to the CoT prompt to enable in-context learning. However, unlike our few-shot prompt, we explain contextual relationships between the five factors and how the factors influence the final classification in addition to providing a positive/negative label, since we want to condition the model to consider contextual relationships between factors and how those factors affect the final classification. We illustrate the full CoT prompt in Fig. 2. Notably, as discussed in II-C, since LVLMs may struggle with interpreting subtle visual factors such as different gestures. To improve performance in this area, we implemented a two-step process. First, GPT-4 was tasked with identifying any gestures present in the image. If no gestures were detected, we introduced a verification component that specifically asked GPT-4 to check for an enumerated list of offensive gestures, including the middle finger, “Loser” symbol, thumbs down, and finger gun. This verification step was only triggered if no gestures were initially detected to avoid limiting the range of potential gestures. As shown in Table II, the verification component significantly improved GPT-4’s accuracy in detecting these explicit gestures, which are strongly correlated with cyberbullying. Although this method enhances accuracy for the most relevant gestures, ideally, a broader range of gestures would also be detected for more comprehensive identification.

IV. Evaluation and Results

A. Experimental Setup

Model. We chose GPT-4 as our evaluation model because it has exhibited strong performance across multiple natural language processing benchmarks, confirming its generalizability and ability to process and understand complex data [33]. Additionally, the model’s large training data size allows it to understand crucial external context in cyberbullying (e.g. Anti-LGBT sentiments.)

Metrics. We opt for a comprehensive approach that employs a set of well-known and academically validated metrics to enable a nuanced understanding of model performance. Specifically, we focus on four key evaluation metrics: accuracy, which provides a general measure of the model’s correct classifications; precision, which assesses the model’s ability to avoid false positives; recall, which evaluates the model’s sensitivity/capability to identify true positives; and the F1 score, which offers a balanced measure of the model’s precision and recall.

Baseline. We use a multimodal model consisting of a linear classifier on features extracted with pretrained models for 4 of the 5 factors (Body Pose, Facial Expression, Gesture, Weapons) fused with general image features extracted with a convolutional neural network that was designed and trained in [9] as a baseline. The model achieved SOTA results, significantly outperforming previous offensive content detectors, such as Google Vision Safe Search [34], Amazon Rekognition Content Moderation [35], Clarifai NSFW [36], Yahoo Open NSFW [37], and DeepAI Offensive Content Detector [38].

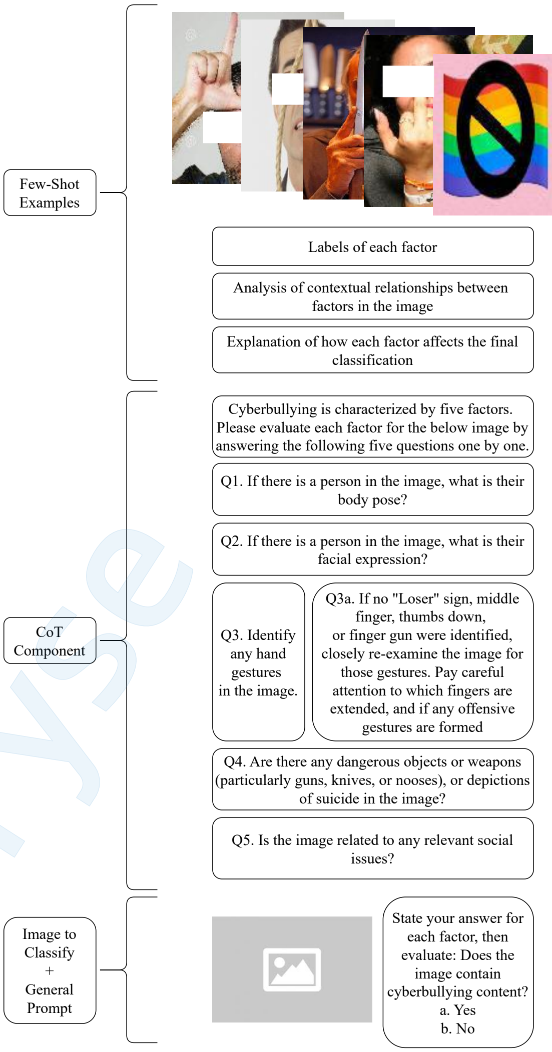

Few-Shot Examples
Labels of each factor
Analysis of contextual relationships between factors in the image
Explanation of how each factor affects the final classification
CoT Component
Cyberbullying is characterized by five factors. Please evaluate each factor for the below image by answering the following five questions one by one.
Q1. If there is a person in the image, what is their body pose?
Q2. If there is a person in the image, what is their facial expression?
Q3. Identify any hand gestures in the image.
Q3a. If no "Loser" sign, middle finger, thumbs down, or finger gun were identified, closely re-examine the image for those gestures. Pay careful attention to which fingers are extended, and if any offensive gestures are formed
Q4. Are there any dangerous objects or weapons (particularly guns, knives, or nooses), or depictions of suicide in the image?
Q5. Is the image related to any relevant social issues?
Image to Classify + General Prompt
State your answer for each factor, then evaluate: Does the image contain cyberbullying content?
a. Yes
b. No

Fig. 2: CoT Prompting Strategy

B. Analysis of Different Prompts

In our experiment, we evaluated four distinct prompting strategies: General Prompt (GP), General Prompt with Cyberbullying Definition (GPwDef), Few-Shot Learning Prompt (Few-Shot), and Chain-of-Thought Reasoning Prompt (CoT). The performance results are summarized in Table III, showing significant variation in the effectiveness of these strategies for guiding GPT-4 in detecting cyberbullying images.

As evidenced by the results in Table III, different prompting strategies exhibit significantly varying levels of effectiveness in guiding GPT-4 for cyberbullying detection. CoT outperforms the other prompts across all metrics, achieving the highest accuracy of 0.95, precision of 0.98, and F1 score of 0.95. This suggests that the task-specific information provided by the

TABLE III: Comparison of different prompts and baseline

Prompt	Accuracy	Precision	Recall	F1
GP	0.50	0.67	0.02	0.04
GPwDef	0.51	0.50	0.00	0.00
Few-Shot	0.55	0.66	0.24	0.36
Baseline [9]	0.93	0.94	**0.97**	0.95
CoT	**0.95**	**0.98**	0.92	**0.95**

CoT prompt, such the factors associated with cyberbullying, is crucial for the model's understanding of cyberbullying.

In comparison, Few-Shot prompting, despite also incorporating task-specific information through labeled examples, falls short with an accuracy of 0.55 and F1 score of 0.36. This contrast highlights the importance of structured reasoning over simple example-based learning. While Few-Shot provides general examples of cyberbullying, it does not guide the model through the critical reasoning process, resulting in lower overall effectiveness. CoT's structured approach, on the other hand, enables the model to process the complex and context-dependent nature of cyberbullying images more effectively. Furthermore, the five factors being explicitly enumerated in the CoT prompt and the higher performance relative to few-shot confirms that accurate evaluation of the five factors is crucial for cyberbullying classification.

Our CoT approach also outperforms the previous SOTA baseline [9] on three of our four metrics. Crucially, the CoT approach achieves this without the need for extensive training datasets or the comprehensive training process required by the baseline model. While the baseline relies on a multimodal model trained on large datasets with specific feature extractors, CoT leverages pre-trained LVLMs, making it a far more efficient and scalable solution. However, the CoT prompt does exhibit a slightly lower recall of 0.92 compared to the baseline's 0.97, indicating that GPT-4 may miss some true positives. This suggests that while CoT excels in reducing false positives, further refinement is needed to improve its sensitivity to subtle instances of cyberbullying.

Interestingly, the general prompt with a definition performs slightly worse than just a general prompt, suggesting that the model's interpretation of the definition of cyberbullying is inconsistent with cyberbullying in real-world images. This further confirms that the task-specific understanding provided by the CoT prompt is crucial for LVLMs in cyberbullying image detection.

V. Related Work

Cyberbullying has become a critical societal issue, receiving extensive attention from researchers across multiple disciplines, including psychology, sociology, and behavioral sciences. Early research primarily focused on the psychological impacts of cyberbullying, exploring how online harassment affects mental health, self-esteem, and well-being, particularly among adolescents [3], [7], [18]. As the internet and social media platforms have evolved, so too has the nature of cyberbullying, prompting increased interest from the computer science community in the development of automated detection and prevention techniques.

Much of the initial work in cyberbullying detection has concentrated on text-based content. Studies in this domain have employed various machine learning and natural language processing techniques to detect harmful messages within textual data on social media platforms. For instance, classical machine learning approaches such as support vector machines (SVMs) and decision trees have been applied to classify text as cyberbullying or non-cyberbullying [39]. More recent studies such as [40] have incorporated deep learning models, leveraging architectures like recurrent neural networks (RNNs) and transformers to capture linguistic patterns indicative of cyberbullying.

With the rise of visual content on platforms, researchers have begun shifting focus towards cyberbullying in images, which presents unique challenges due to the complexity and context-dependent nature of visual cues. Recent studies have explored methods for detecting cyberbullying in multimedia content. In Almomani et al. [41], transfer learning was employed by applying a classification head to pretrained convolutional neural network (CNN) backbones. This approach showed promising results, particularly when dealing with small datasets, although its scalability to larger, more complex datasets remains uncertain. Building on this, Vishwamitra et al. [9] collected a large-scale dataset of cyberbullying images and introduced a multimodal model that integrates features from multiple visual factors such as body pose, facial expression, and gestures. Their model combined these feature representations with an end-to-end trained CNN feature extractor, achieving SOTA performance in cyberbullying detection. This multimodal approach effectively demonstrated the importance of considering multiple visual elements in combination. In our work, we investigate the capabilities of pretrained LVLMs in cyberbullying image detection. Through our CoT-based approach, we match or exceed the performance of previous approaches without the need for a training process.

VI. Conclusion and Future Work

In this paper, we investigated the application of Large Vision-Language Models (LVLMs) for cyberbullying image detection, specifically focusing on their ability to understand and evaluate key visual factors such as body pose, facial expression, gestures, objects, and social context. By employing a Chain-of-Thought (CoT) prompting strategy, we demonstrated that task-specific reasoning pathways significantly enhance the model's accuracy and precision, outperforming both traditional few-shot methods and the state-of-the-art (SOTA) detection approach. Additionally, unlike traditional methods that require extensive training datasets and complex training processes, our CoT-based strategy leverages pre-trained LVLMs, reducing the need for large-scale labeled data and computational resources. Despite this, our method matches or exceeds the performance of models that rely on comprehensive training.

Despite these advancements, our approach still faces challenges, particularly in improving recall for subtle forms of cyberbullying. Further refinements, such as enhancing sensitivity to these subtleties, may help address this issue in future work. Exploring additional prompting strategies or fine-tuning LVLMs specifically for cyberbullying detection could further improve performance.

In conclusion, our findings contribute to the growing field of cyberbullying detection by demonstrating the potential of LVLMs combined with task-specific reasoning strategies. As LVLMs continue to evolve, our methodology will serve as important foundations for the development of more reliable and effective systems to combat online harassment and promote safer digital environments.

ACKNOWLEDGEMENTS

This material is based upon work supported in part by the National Science Foundation (NSF) under Grant No. 2120369, 2129164, 2228617, 2245983, and 2329704.

REFERENCES

[1] Pew Research Center. A majority of teens have experienced some form of cyberbullying, 2018. Accessed: 2024-09-09.

[2] National Center for Education Statistics. Student reports of cyberbullying: Results from the 2019 school crime supplement to the national crime victimization survey, 2022. Accessed: 2024-09-09.

[3] Robin M. Kowalski, Susan P. Limber, and Annie McCord. A developmental approach to cyberbullying: Prevalence and protective factors. *Aggression and Violent Behavior*, 45:20–32, 2019. Bullying and cyberbullying: Protective factors and effective interventions.

[4] Instagram, 2024. Accessed: 2024-09-09.

[5] Snapchat, 2024. Accessed: 2024-09-09.

[6] Tiktok, 2024. Accessed: 2024-09-09.

[7] Sara Bastiaensens, Heidi Vandebosch, Karolien Poels, Katrien Van Cleemput, Ann DeSmet, and Ilse De Bourdeaudhuij. Cyberbullying on social network sites: An experimental study into bystanders' behavioral intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, 31:259–271, 2014.

[8] Zahra Ashktorab. A study of cyberbullying detection and mitigation on instagram. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, pages 126–130, 2016.

[9] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. Towards understanding and detecting cyberbullying in real-world images. *Proceedings 2021 Network and Distributed System Security Symposium*, 2021.

[10] Nureni Azeez, Sunday Idiakose, Chinazo Onyema, and Charles van der Vyver. Cyberbullying detection in social networks: Artificial intelligence approach. *Journal of Cyber Security and Mobility*, 10:745–774, 06 2021.

[11] Abdulsamad Al-Marghilani. Artificial intelligence-enabled cyberbullying-free online social networks in smart cities. *International Journal of Computational Intelligence Systems*, 15(1):9, 2022.

[12] Celestine Iwendi, Gautam Srivastava, Suleman Khan, and Praveen Kumar Reddy Maddikunta. Cyberbullying detection solutions based on deep learning architectures. *Multimedia Systems*, 29(3):1839–1852, 2023.

[13] Aaminah Ali and Adeel M Syed. Cyberbullying detection using machine learning. *Pakistan Journal of Engineering and Technology*, 3(2):45–50, 2020.

[14] B Sri Nandhini and JI Sheeba. Cyberbullying detection and classification using information retrieval algorithm. In *Proceedings of the 2015 international conference on advanced research in computer science engineering & technology (ICARCSET 2015)*, pages 1–5, 2015.

[15] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[16] Nishant Vishwamitra, Keyan Guo, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, and Hongxin Hu. Moderating new waves of online hate with chain-of-thought reasoning in large language models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 788–806, 2024.

[17] Keyan Guo, Ayush Utkarsh, Wenbo Ding, Isabelle Ondracek, Ziming Zhao, Guo Freeman, Nishant Vishwamitra, and Hongxin Hu. Moderating illicit online image promotion for unsafe user generated content games using large Vision-Language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 5787–5804, Philadelphia, PA, August 2024. USENIX Association.

[18] Justin W. Patchin and Sameer Hinduja. Cyberbullying and self-esteem. *Journal of School Health*, 80(12):614–621, 2010.

[19] Google search, 2024. Accessed: 2024-09-09.

[20] Bing search, 2024. Accessed: 2024-09-09.

[21] Baidu search, 2024. Accessed: 2024-09-09.

[22] Flickr, 2024. Accessed: 2024-09-09.

[23] X (formerly twitter), 2024. Accessed: 2024-09-09.

[24] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023. Accessed: 2024-09-09.

[25] Cyberbullying definition. https://www.pacer.org/bullying/info/cyberbullying/, 2021.

[26] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[27] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[28] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[29] Robert L Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. Cutting down on prompts and parameters: Simple few-shot learning with language models. *arXiv preprint arXiv:2106.13353*, 2021.

[30] Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573, 2023.

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[32] Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-Thought Hub: A Continuous Effort to Measure Large Language Models' Reasoning Performance. *arXiv preprint arXiv:2305.17306*, 2023.

[33] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[34] Google Cloud. Cloud vision safesearch detection, 2024. Accessed: 2024-09-09.

[35] Amazon Web Services. Amazon rekognition content moderation, 2024. Accessed: 2024-09-09.

[36] Clarifai. Clarifai nsfw model, 2024. Accessed: 2024-09-09.

[37] Yahoo. Yahoo open nsfw model, 2024. Accessed: 2024-09-09.

[38] DeepAI. Offensive content detection api, 2024. Accessed: 2024-09-09.

[39] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, volume 2, pages 241–244, 2011.

[40] Apoorva K G and D Uma. Detection of cyberbullying using machine learning and deep learning algorithms. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–7, 2022.

[41] Ammar Almomani, Khalid Nahar, Mohammad Alauthman, Mohammed Azmi Al-Betar, Qussai Yaseen, and Brij B. Gupta. Image cyberbullying detection and recognition using transfer deep machine learning. *International Journal of Cognitive Computing in Engineering*, 5:14–26, 2024.