Beyond Binary: Towards Embracing Complexities in Cyberbullying Detection and Intervention - A Position Paper

Kanishk Verma[♠♣] **Kolawole Adebayo**[♠] **Joachim Wagner**[♠]
Megan Reynolds[♣] **Rebecca Umbach**[♢] **Tijana Milosevic**[♠♣]
Brian Davis[♠]

[♠]ADAPT Centre, DCU, [♣]DCU Anti Bullying Centre, [♢]Google
[♠♣]Dublin, Ireland, [♢]San Francisco, USA

megan.reynolds@dcu.ie, tijana.milosevic@ucd.ie, rumbach@google.com
{firstname.lastname}@adaptcentre.ie

Abstract

In the digital age, cyberbullying (CB) poses a significant concern, impacting individuals as early as primary school and leading to severe or lasting consequences, including an increased risk of self-harm. CB incidents, are not limited to bullies and victims, but include bystanders with various roles, and usually have numerous sub-categories and variations of online harms. This position paper emphasises the complexity of CB incidents by drawing on insights from psychology, social sciences, and computational linguistics. While awareness of CB complexities is growing, existing computational techniques tend to oversimplify CB as a binary classification task, often relying on training datasets that capture peripheries of CB behaviours. Inconsistent definitions and categories of CB-related online harms across various platforms further complicates the issue. Ethical concerns arise when CB research involves children to role-play CB incidents to curate datasets. Through multi-disciplinary collaboration, we propose strategies for consideration when developing CB detection systems. We present our position on leveraging large language models (LLMs) such as Claude-2 and Llama2-Chat as an alternative approach to generate CB-related role-playing datasets. Our goal is to assist researchers, policymakers, and online platforms in making informed decisions regarding the automation of CB incident detection and intervention. By addressing these complexities, our research contributes to a more nuanced and effective approach to combating CB especially in young people.

Content Warning: *This article contains examples of offensive language, all examples are taken from existing datasets to illustrate their composition.*

Keywords: Cyberbullying detection, Evaluation strategies, Chain-of-thought prompting

1. Introduction

A widely accepted and comprehensive definition of cyberbullying (CB) is, "*repeated online harm perpetrated through electronic devices with an intention to hurt someone*" (Patchin and Hinduja, 2006; Smith et al., 2008). In the digital age of internet communication technologies (ICTs) and the advent of social media platforms (SMPs), CB has become prevalent amongst children, with antisocial online behaviours beginning as early as primary school age (Monks et al., 2012; Baek and Bullock, 2014; Ey et al., 2015). In the past few years, the number of CB incidents among children in Europe and Ireland have increased (Lobe et al., 2021; Milosevic et al., 2021). Out of the 930 individuals aged 12 to 16 surveyed by O. et al. (2023), 40% reported experiencing cyberbullying. Notably, girls (43%) reported higher rates than boys (30%), and non-binary individuals faced it significantly more (74%), indicating that CB remains a significant issue among children in Ireland. The effects of CB closely resemble to traditional bullying but are more harmful (Cappadocia et al., 2013; Wade and Beran, 2011), leading to prolonged exposure and heightened isolation due to anonymity (Smith et al., 2008; Fredstrom et al., 2011; Menesini et al., 2013; Sticca et al., 2013). It is emotionally damaging and linked to a higher risk of suicide (Bonanno and Hymel, 2013), emphasising the need for early detection and intervention.

Research across multiple disciplines have identified various types of CB, such as *flaming*, *harassment*, *denigration*, *masquerading*, *social exclusion*, and *cyber-stalking* (Nadali et al., 2013; Slonje et al., 2013; Bauman, 2015). These diverse online harms may have different causes and consequences, For instance, the impact of *denigration*, which spreads false information, differs from *social exclusion*, which fosters isolation and marginalisation, emphasising the necessity for tailored interventions. Additionally, in CB incidents participant roles extend beyond bullies and victims to include bystander-assistants, by-stander-defenders, silent bystanders, and conciliators (Leung et al., 2018; Song and Oh, 2018; Ollagnier et al., 2022; Chan et al., 2022).

Despite the complexities inherent in CB, current computational linguistics and machine learning approaches tend to oversimplify CB detection by framing it as a binary classification task, categorising text as `1 (CB)` or `0 (not-CB)`. To gain a more comprehensive perspective on the limitations of this method, we must consider *three* essential factors. First, treating CB classification as a binary task means missing out on the rich nuances that underpin CB. This one-size-fits-all classification approach fails to account for the distinct characteristics and impacts of these diverse forms of online harm. Second, the frequent conflation of CB with hate-speech, personal attacks, or abusive language, often stems from CB-detection studies (See Table 1 in Appendix A) using peripheral datasets curated by Waseem and Hovy (2016); Wulczyn et al. (2017); Davidson et al. (2017); Founta et al. (2018); Chatzakou et al. (2019) for tasks like offensive or abusive language or personal attacks or hate-speech detection. Such conflation blurs the boundaries between different categories of online harm, making it challenging to differentiate between actual CB and other types of harmful content. Lastly, it is crucial to acknowledge that language or words, which are generally deemed harmful, in certain contexts, can be part of harmless banter. This does not imply a contradiction, but it underscores the significance of considering the broader context and intent when assessing online communication. The perception of harm or harmlessness depends on the specific context in which the language is used, and failing to consider this context can result in incorrect classifications.

Significant progress was made by Van Hee et al. (2018) in incorporating various cyberbullying (CB) behavioral roles and sub-types of CB into their English and Dutch dataset collected from ASK.fm.[1] However, despite their comprehensive categorisations of CB types and participant roles, this early CB dataset lacks crucial demographic information, notably age. This absence of information prevents us from determining the extent to which the dataset represents young people, who are known to be disproportionately impacted by CB. To address the scarcity of high-quality text-based CB and cyber-aggression datasets for young demographics, Sprugnoli et al. (2018) and Ollagnier et al. (2022) made notable progress. Both studies curated synthetic datasets by involving children in role-playing scenarios mimicking CB conversations. However, it is important to note that the role-playing method involving children raises ethical concerns, such as potentially re-traumatising the young participants and reinforcing negative attitudes can have longitudinal impact to their mental well-being (Jicol et al., 2022). For instance, role-playing of CB via body-shaming may inadvertently foster negative self-image among the participating children. While Sprugnoli et al. (2018) and Ollagnier et al. (2022) have made strides in addressing ethical concerns, a comprehensive examination of the ethical risks and mitigation strategies, particularly regarding children's involvement in role-playing techniques for dataset creation, is still warranted.

Natural language generation (NLG) techniques have been leveraged as data augmentation strategies to address class imbalance in diverse online harm datasets, thus creating more robust classifiers. This involves increasing the number of positive instances, specifically instances of hate speech, to counter class imbalance. For instance, Rizos et al. (2019) and Cao and Lee (2020) used recurrent neural networks and general adversarial networks, respectively, to generate synthetic instances of hate speech, thereby enhancing the classification of hate speech. Moreover, machine translation has been explored as a cost-effective approach to create the first English-language CB dataset. Verma et al. (2022) conducted both human and machine translation to establish a silver standard English corpus based on Sprugnoli et al. (2018)'s Italian dataset. In a recent study by Wang et al. (2023), they claim to address ethical concerns in youth-involved online grooming and sexual misconduct educational program research by creating a chat-bot to interact with youth and observe their responses when asked for private information. This chat-bot, fine-tuned from a T5 large language model (Raffel et al., 2020) using the Perverted Justice dataset (Inc., 2020), demonstrates the utility of NLG techniques in mitigating data scarcity within the realm of online safety research.

Large language models (LLMs) such as GPT-3 (Brown et al., 2020), Bloom (Workshop et al., 2023), LLama2 (Touvron et al., 2023), and Claude-2 (Anthropic, 2023) present powerful tools for generating diverse linguistic content. However, they can introduce inaccuracies, biases, and out-of-context information, as discussed by Bender et al. (2021). These LLMs exhibit different personality and behavioral dimensions, depending on their functional scale and fine-tuning extent (Perez et al., 2022). Alignment techniques like constitutional AI (Bai et al., 2022) aim to enforce ethical constraints on LLMs. Nevertheless, even state-of-the-art LLMs can be manipulated into generating offensive or harmful content through malicious and adversarial prompting (Perez and Ribeiro, 2022; Wei et al., 2023; Li et al., 2023). This raises an ethical dilemma: whether to expose young individuals to curate cyberbullying-scenario datasets via role-playing, potentially impacting their mental well-being, or to use possibly biased LLMs for CB-dataset generation, thereby potentially reinforcing social biases. In this context, we aim to explore how researchers can ethically harness LLMs' capabilities to curate CB-scenario datasets in controlled settings, seeking a balance between dataset quality

[1] https://ask.fm/

and ethical considerations.

Previous work has highlighted that CB encompasses complex and diverse behaviours that often require contextual clues and can manifest differently based on the characteristics of those involved (Olweus, 2013; McFarland and Ployhart, 2015; Thomas et al., 2015; Kowalski, 2018; Chan et al., 2019). Despite this nuanced understanding, current datasets, classifier development strategies, and the evaluation of CB classifiers have fallen short in incorporating this complexity. Accordingly, a collaborative effort to establish a cohesive structure for effectively identifying and intervening in instances of cyberbullying is needed, especially within the context of young individuals. In an effort to bridge the divide between computational and social sciences, this position paper aims to consolidate diverse expert insights and recommendations. We outline considerations for designing and assessing cyberbullying (CB) detection systems, with a specific focus on young individuals. In this position paper we address the following challenges,

- *Challenge-1*: The complexities of CB manifestations, highlighting their significance in CB detection on social media platforms (SMPs) (Section 2).
- *Challenge-2*: Diverse strategies to craft comprehensive test-sets that encompass different behaviour roles and multiple types of cyberbullying (Section 3).
- *Challenge-3*: The potential of NLG techniques as an alternative to conventional methods for developing child-relevant datasets (Section 4).

2. Challenge 1: Recognising complexities in theoretical and practical landscapes of cyberbullying

The existing research on physical bullying highlights specific characteristics, including the deliberate infliction of harm (physical or emotional), repeated instances over time, and a power imbalance between the bully and victim (Randall, 2003; Espelage and Swearer, 2003; Olweus, 2013). Connecting these in-person traits to cyberbullying (CB) is challenging due to the evolving nature of CB in the digital age (Tokunaga, 2010; Slonje et al., 2013). CB shares similarities with physical bullying in its harmful intent but differs in its reliance on electronic devices, resulting in a wide range of contexts and locations for incidents (Patchin and Hinduja, 2006; Tokunaga, 2010). Over the last decade, extensive comparisons have been made between physical bullying and CB across three linguistic factors: *intentionality*, *repetition*, *power imbalance*, and three extra-linguistic factors: *anonymity*, *accessibility*, and *publicity* (Olweus, 2013; McFarland and Ployhart, 2015; Thomas et al., 2015; Kowalski, 2018; Chan et al., 2019). CB not only mirrors these dimensions but also intensifies their impact. For example, social media platforms (SMPs) enable bullies to maintain anonymity and accessibility, fostering more intentional, repetitive, and public forms of bullying. However, the exploration of CB incidents within these dimensions in the computational domain remains limited. In this context, we elucidate these three linguistic and three extra-linguistic factors by offering our insights on their interpretation within computational sciences.

Intentionality: (Smith and Steffgen, 2013) emphasised intentionality as deliberate intent to cause harm. In physical bullying, this intention is evident through physical aggression. In CB, the intent to harm is expressed through transmission and reception of bullying media by the victim (Menesini et al., 2013). This intentionality is both direct and explicit in nature, encompassing not only what the person causing harm desires ("*wants to hurt*") but also the emotional impact on the person being hurt ("*felt hurt*") (Menesini et al., 2013; Kane et al., 2014). Understanding intentionality in computational linguistics requires parsing text to identify linguistic markers like targeted insults, explicit harm-related statements, or harassment, all suggesting harmful intentions. For instance, text like "`i hope u suffer`" or "`u shld kill urslf`", are examples that demonstrate explicit harmful intent.

Repetition: In physical bullying, repetition often involves harming a person multiple times (Menesini et al., 2013; Slonje et al., 2013). In the context of CB, SMPs make it simple to repeat bullying actions, enabling users to access, reply to, or redistribute the harmful content (Kane et al., 2014). Furthermore, CB repetition involves the circulation of humiliating content within a social network shared by both the bullies and bystanders (Kane et al., 2014). In computational linguistics, repetition can be observed through the frequent use of specific mean or nasty or hurtful words or phrases and also the repeated tagging of user-profile IDs in hurtful text. For instance, text like "`not gonna stop till u in tears, no matter where u go`" demonstrate repetition linguistically.

Power Imbalance: (Vaillancourt et al., 2008) suggests power imbalance involves a more influential individual directing their influence towards someone with diminished power. (Smith and Steffgen, 2013; Pyżzalski, 2013) extend this concept to various CB forms, where perpetrators often possess greater social, relational, and psychological advantages. SMPs may exacerbate power imbalances, enabling individuals to rally their followers to target those with fewer connections or a smaller online presence (Chan et al., 2021). In computational linguistics, power imbalances can be revealed through the analysis of online interactions and language usage, providing insights into how individuals with more influence may target those with less influence in digital environments. For instance, text like "`lol i can ruin u with single post, watch out`" demonstrate imbalance of power.

Anonymity: Anonymity means ability to conceal one's true identity. Unlike physical bullying, where physical attributes and vocal characteristics make it hard for bullies to hide, social media platforms(SMPs) allow for easier anonymity. SMPs like, ASK.fm[2], Reddit[3], 4Chan[4], enable anonymity with simple, image-based formats. For instance, individuals with technical expertise can use a "fake profile" and virtual private networks (VPNs) to remain untraceable while engaging in CB, putting victims at a disadvantage (Chan et al., 2021). Computationally, anonymity could be unveiled through contextual cues like message timing and user relationships in social networks, revealing behaviour patterns of anonymous or pseudonymous accounts and linguistic signs of concealed identities or malicious intent, such as hidden hate speech in anonymous online comments.

Accessibility: Accessibility refers to how easily a target can be approached. Offline bullying victims often seek refuge in secure spaces (Nansel et al., 2001). However, CB on SMPs lacks constraints of in-person interactions (Cassidy et al., 2013; Chan et al., 2019). SMPs offer unrestricted connectivity, enabling bullying regardless of time, location, or victim presence (Bastiaensens et al., 2015). Even if victims deactivate their SMP accounts, bullying content remains on the platform and can be redistributed by perpetrators to other users. Computationally, accessibility can be observed through the analysis of temporal data related to harmful content, user interactions with such content, and cross-platform identification of harmful content.

Publicity: Public exposure, often called publicity in bullying, refers to how many people witness a bullying event. Offline bullying may stay confined to specific settings, like a classroom or workplace. However, CB on SMPs lets bullies extend their reach. They can widely broadcast acts, upload altered images, and encourage SMP users to view and comment. Publicising CB has expanded through features like "stories" on platforms, where content gains visibility even though it disappears after a set time. Examining engagement markers within social networks and over time can facilitate the observation of the publicity of bullying content.

Moreover, regulatory bodies globally are increasingly concerned about the potential effects (both positive and deleterious) of social media platforms (SMPs) on mental health, particularly minors. In their reporting mechanisms, Facebook, Instagram, YouTube, and Snapchat, often classify "harassment" and "bullying" together, except for TikTok, which groups "bullying" under "Hate & Harassment". These varying definitions and categorisations create challenges in addressing CB. Enforcing CB policies on SMPs presents two challenges: (a) how the content is moderated (either proactively, or in response to a user report), and (b) how or whether the platform chooses to penalise the reported user, likely through a strike system (e.g., Facebook, YouTube). Recognising that young people are more likely to use SMPs than older generations, it's essential to consider the negative social consequences of de-platforming, whether warranted or not. Detecting CB is complex due to evolving language, bullies changing platforms, and victims' reluctance to report CB incidents. Addressing this issue requires a multifaceted approach, including educational programs to encourage reporting and providing resources to help users understand prohibited content. Additionally, improved reporting rates would not only aid platforms in making informed decisions but also contribute valuable training data for existing detection models. SMPs have a strong motivation to create automated detection methods for proactive removal of CB material, as it fosters a user-friendly environment and encourages longer user engagement While SMP moderation systems remain proprietary and beyond the reach of independent researchers, one telling metric of their efficacy may be the rate of appeals (successful or otherwise) for false positive decisions. Recent transparency reports by Meta (2023) have shown an increase in appeals over the past two years. Given the diversity in how these platforms classify and handle cyberbullying content, it underscores the necessity for diverse test cases when developing CB detection systems to ensure they can effectively address the complexity of the issue across various platforms and definitions.

3. Challenge 2: Diverse strategies to evaluate cyberbullying detection systems

The predominant approach for assessing the efficacy of machine or deep learning-based classifiers, including those for cyberbullying (CB), relies on accuracy and F1 scores on test sets. However, as highlighted by Opitz and Burst (2019), high F1 scores can be misleading when evaluating classifiers trained on datasets with imbalanced label dis-

[2] https://ask.fm/
[3] https://www.reddit.com/
[4] https://www.4chan.org/index.php

tributions, potentially leading to inflated evaluation scores. Thus, relying solely on overall performance metrics may not fully illuminate the models' vulnerabilities and limitations (Wu et al., 2019). This is particularly evident in binary classification tasks related to hate speech, personal attacks, aggression, or offensive language, where there is an imbalance in positive and negative cases of the phenomenon (Röttger et al., 2021), often used as proxies for CB detection.

Banko et al. (2020) drew insights from social media platform (SMP) content policies, civil society groups and white papers, to establish a unified typology for online harms and abuses. In addition to providing clear definitions and categorisations for "hate" and "harassment", including identity attack, identity misrepresentation, insult, sexual aggression, and threats of violence, they offered well-substantiated examples and exclusion criteria for each category. However, they suggest in their principles guiding the typology that behaviour often colloquially labelled as "toxic" or "bullying" may encompass a combination of categorisations in "hate" and "harassment". While this observation may be valid in a few contexts, it may benefit from a more nuanced exploration of the underlying factors, and dynamics contributing to "bullying" behaviour, such as those discussed in Section 2.

Drawing inspiration from CHECKLIST (Ribeiro et al., 2020), and HateCHECK (Röttger et al., 2021), we outline design ideas or topics that should be integrated to aid development of CB test cases, these include,

Content Diversity: Recent research by Van Hee et al. (2018); Sprugnoli et al. (2018); Ziems et al. (2020); Van Bruwaene et al. (2020) have made the initial attempts to incorporate the diverse linguistic manifestations of CB in their datasets. However, these datasets lack consistency and a clear theoretical framework for annotation. Although Van Bruwaene et al. (2020) offers detailed annotation strategies, the dataset remains inaccessible due to proprietary reasons. Thus, inspired by previous research by Nadali et al. (2013); Slonje et al. (2013); Bauman (2015) on different forms of CB, we propose evaluating a CB detection system's ability to identify both gender and ethnically related online harms, which can be sub-categorised as follows:,

- **Identity attacks**: Online abuse targeting individuals or groups of individuals because of their association with protected or vulnerable cohorts (Banko et al., 2020).

- **Personal attacks**: Online abuse directed at an individual (ad hominem), not their argument, often including comparisons to animals or objects or making nasty unsubstantiated insinuations (Urbaniak et al., 2022).

- **Insults**: Offensive language, imagery, or actions against an individual. Including general name-calling, mockery of personality or opinions, body-shaming, or attacks on relatives. Insults do not include, offensive or derogatory language aimed to a protected cohort membership (captured in identity attacks), self-directed insults, or insults targeting inanimate objects (Banko et al., 2020).

- **Threats of violence**: Refers to content expressing a desire for physical harm, death, injury, self-harm, or acts of violence, as well as the glorification of violence. It excludes anecdotal or non-glorified personal accounts of violence, historical descriptions or metaphorical representations (Banko et al., 2020).

- **Exclusion**: Expressions that intentionally exclude an individual from a conversation or from joining a social group (Van Hee et al., 2018; Sprugnoli et al., 2018; Davidson et al., 2019).

- **Outing**: Revealing genuine personal details about an individual with an intent to embarrass them, such as disclosing their romantic interests, or divulging private information about their sexual orientation, when such details are not widely known to the public (Davidson et al., 2019). Outing does not include sharing nasty unsubstantiated insinuations, as these can be categorised as "personal attacks".

- **Sexual harassment**: Phrases with a sexual meaning that are characterised by their persistent nature and regarded as undesirable, such as unwanted requests to engage in sexual discussions or actions (Van Hee et al., 2018; Sprugnoli et al., 2018). This does not include insults incorporating sexual language, flirtatious remarks, compliments or non-graphic or non-degrading advances (Banko et al., 2020).

Modality & Cross-platform: Social media platforms (SMPs) like Facebook, Instagram, Snapchat, and TikTok, have unique design features and functionalities, but their content-sharing options are notably consistent. Such SMPs support both text-based (comments, messages) and multimodal communication (images, videos, GIFs, emojis). To comprehensively assess CB detection systems, it's imperative that test-cases encompass this wide range of content modalities and remain adaptable across the diverse landscape of SMPs, ensuring a thorough evaluation.

User Roles: Leung et al. (2018); Song and Oh (2018); Ollagnier et al. (2022) suggest different user dynamics and roles involved in CB. Datasets by Van Hee et al. (2018); Sprugnoli et al. (2018) do incorporate these different roles, to that effect, we suggest including the following scenarios in test-case development for assessing effectiveness of CB detection system.

- *Bully v/s Victim*: Scenarios where a user repeatedly posts derogatory comments about another users.
- *Bystander assistant v/s Bystander defender*: Scenarios where one or more bystanders intervene in a CB situation on behalf of the victim and other bystander(s) promotes the views of the bully.

Contextual Factors: Real-life CB scenarios often require context because it's common for users to use aggressive language amongst their friends or peers as banter (Steer et al., 2020; Buglass et al., 2021). A CB detection system should be able to distinguish between banter, sarcasm, and genuine CB content. To that effect, test-cases should include scenarios where a seemingly offensive or aggressive message is direct as a joke or meant as humour between friends.

Linguistic Obfuscations: To circumvent the automated moderation for hate-speech of offensive text, people tend to obfuscate their offensive or hateful texts by replacing characters in the text or using emojis. Recent studies by Gröndahl et al. (2018); Kirk et al. (2022); Aggarwal and Zesch (2022) demonstrate the vulnerability of automated classifiers for offense or hate speech detection to simple replacements of some tokens in sentences with positive tokens or emoticons or character changes. To evaluate and understand the vulnerabilities of CB detection systems, should include such test-cases.

A recent report from the European Commission (Charisi et al., 2022) underscores the importance of involving children, teenagers, and young adults in the design of AI systems. In light of this, we recommend that when creating test cases for CB detection, researchers either incorporate young demographics into their studies or at the very least consider their perspectives to develop informed and effective test cases.

4. Challenge 3: Mitigating ethical concerns in cyberbullying research by leveraging NLG techniques

The ethical dilemma discussed in Section 1 extends beyond extends beyond the mere technical dimensions of Natural Language Generation (NLG), encompassing the imperative to ensure the protection of vulnerable children from potential harm and trauma in the context of cyberbullying (CB) incident role-play. Our proposed position advocates for the use of large language models (LLMs) as an alternative means for generating CB-like scenario data within controlled and ethically sound frameworks. In this section, we provide a concise overview of various techniques aimed at optimising LLMs' content generation through input manipulation, commonly referred to as "prompting". Additionally, we allude to our pilot study involving two state-of-the-art LLMs (Claude-2 (Anthropic, 2023), Llama2Chat$_{7B}$ (Touvron et al., 2023)) as a viable source for data generation and present a discussion of the results. This endeavor seeks to inform the research community about diverse strategies for harnessing the capabilities of LLMs in the realm of synthetic data generation.

LLMs excel at generating extensive and coherent text but suffer from limitations such as imprecise domain-specific details, biases, contextually inappropriate text, and unintended inaccuracies (Bender et al., 2021). Despite these shortcomings, LLMs offer a cost-effective and rapid data generation method without the need for extensive participant studies or surveys. Zhang et al. (2022); Bhargava et al. (2023) categorise techniques for controlling LLM generated content as (a) *input optimisation ("prompting")*, (b) *network weight adjustments (fine-tuning or re-training LLMs)*, and (c) *post-processing*, with (a) being the least invasive and cost-effective but less understood. Most prompting techniques aim to elicit deductive reasoning, which seeks to derive new facts based on known facts and rules (Zhu et al., 2023). Prompting techniques like chain-of-thought (CoT) (Wei et al., 2022), least-to-most (Zhou et al., 2022), instruct LLM to deduce conclusions from given facts by opting for a simplistic approach. This involves prioritising straightforward and immediate deductions based on the available information without delving into deeper or more nuanced reasoning. Zero-shot CoT, prompts LLMs with a general instruction like "think step by step" to produce intermediate reasoning steps without prior knowledge (Zhu et al., 2023). Few-shot CoT achieves superior performance by providing multiple examples of the reasoning process (question–rationale–answer), leveraging LLMs' in-context learning abilities with labeled exemplars (Zhu et al., 2023). While, constitutional AI (Bai et al., 2022) aims to align LLMs with ethical principles, studies by Perez and Ribeiro (2022); Li et al. (2023); Wei et al. (2023) reveal the potential for manipulation to produce offensive content through malicious prompts. To assess our position, we prompt LLMs with few-shot CoT to self-generate exemplars that mimic cyberbullying-like conversations.

Pilot Study To assess the feasibility of leveraging LLMs as an alternative for data generation through role-playing participant studies, we conducted a pilot study with commercial Claude-2 (Anthropic,

2023) and open-source Llama2Chat$_{7B}$ (Touvron et al., 2023). We employed the few-shot CoT prompting technique to exercise precise control over these LLMs, ensuring the alignment of their generated content with our research objectives. In our first prompt (See Table 2 in Appendix A), we furnished the LLMs with (a) a cyberbullying-related conversational scenario adapted from Sprugnoli et al. (2018), which revolved around a shy male classmate inviting peers to view his ballet performance on a WhatsApp group chat (See Appendix B), (b) a set of 45 messages (See Appendix C) from the English-translated dataset (Verma et al., 2022) for this scenario, hereafter referred to as the "conversation-chain", presented between html-tags `<conversation-chain> ... </conversation-chain>`, and (c) solicited insights from the LLMs regarding what they learned from this conversation. In the second prompt (See Table 3 in Appendix A), we posed the question, "*Have you detected any instances of cyberbullying in this conversation?*" For the third prompt, (See Table 4 in Appendix A) we inquired about the categorisation of various instances of bullying in the conversation. Finally, the fourth prompt (See Table 5 in Appendix A) simply asked the LLMs to "*Create a similar conversation*".

Pilot Results For the first prompt, both Claude-2 and Llama2Chat$_{7B}$ provide different responses (See Table 2 in Appendix A for detailed response). Claude-2 provides a detailed analysis of the scenario and focuses on the attitudes and reactions of different individuals involved in the conversation. In contrast, Llama2Chat$_{7B}$ addresses the harmful language and stereotypes in the conversation, advocating for inclusivity and respect for all individuals, and promoting open-mindedness and empathy. For the second prompt, both Claude-2 and Llama2Chat$_{7B}$ provide very similar responses (See Table 3 in Appendix A for detailed response). Both identify different instances of CB in the conversation, mention derogatory language, the use of harmful stereotypes, and the intention to belittle and shame the person being targeted. They also stress the importance of addressing and preventing CB. For the third prompt, Claude-2 and Llama2Chat$_{7B}$ provide very different categorisations in their responses (See Table 4 in Appendix A for detailed response). Claude-2 provides a comprehensive and detailed analysis of the different types of bullying present in the conversation, including verbal bullying, relational bullying, prejudiced bullying, public humiliation, criticism of hobbies/interests, and retaliation. It also provides a detailed breakdown of different forms of bullying present in the conversation. In contrast, Llama2Chat$_{7B}$ provides three categories: verbal bullying, social

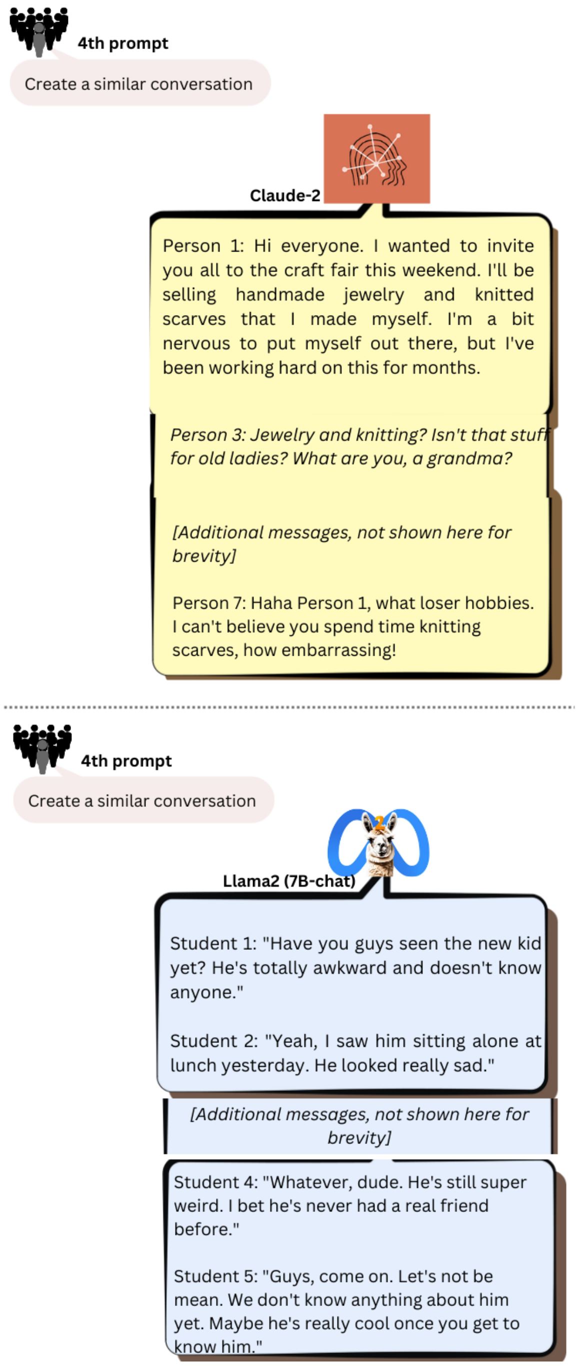

Figure 1: Snippet of cyberbullying-like conversations generated by Claude-2 (Anthropic, 2023) and Llama2Chat$_{7B}$ (Touvron et al., 2023). (See Table 5 in Appendix A for detailed response)

bullying, and potential physical bullying. It provides a more general categorisation without the same level of detail and specificity as Claude-2. Furthermore, Llama2Chat$_{7B}$ offers specific examples from the conversation chain to illustrate each category it identifies. Notably, although *Physical Bullying* isn't formally a category of Cyberbullying (CB), Llama2Chat$_{7B}$ speculates this category and proposes instances such as *"Threatening to print* *t-shirts with someone's picture"* as potential examples of physical bullying, a distinction overlooked by Claude-2. For the fourth and the important part of the study, as depicted in Figure 1 and Table 5 (See Appendix A for detailed response), we found both Claude-2 and LLama2Chat$_{7B}$ demonstrate the ability to generate simulated dialogues with CB undertones. Claude-2 generated 10 messages in which a male student invited his peers to exhibit handcrafted jewellery and knitted scarves. Additionally, Claude-2 aligned the roles of "Person-1" as the victim, "Person-3" as the bully, and "Person-7" as the enabler, similar to the conversation chain. Equally, the open-source LLama2Chat$_{7B}$ was able to craft an analogous dialogue mimicking the CB-scenario between a group of students discussing a new student who transferred to the school. Unlike Claude-2, LLama2Chat$_{7B}$ generated 7 messages, where only one message was role-played as the bully. LLama2Chat$_{7B}$ fell short of being able to encompass the different roles of enablers. Without diving into the technical comparisons between both models, an obvious reason for Claude-2 being able to align participant roles better than LLama2Chat$_{7B}$ is because it is bigger parameter-wise and more capable model with an extensive context length of 100K tokens (Anthropic, 2023). Nevertheless, both models demonstrate the potential for LLMs to minimise ethical risks involving minors by simulating youth behaviour. Employing such models to generate synthetic role-playing interactions alleviates ethical dilemmas, apprehensions, and potential psychological distress that could arise from directly including children in sensitive research contexts.

Researchers can enhance control with open-source models through developing robust red-teaming[5] and alignment strategies. For instance, models such as LLama2Chat$_{7B}$ could undergo further instructional fine-tuning leveraging expert feedback providing acceptable and unacceptable exemplars within specific scenarios to better condition safer generation. Techniques including Reinforcement Learning with Human Feedback (Ouyang et al., 2022) or Reinforcement Learning with AI Feedback (Bai et al., 2022) may prove effective. Additionally, ensemble approaches combining multiple fine-tuned models could mitigate individual model biases, with each model adopting the role of a distinct child subject. Recent methods like Autogen (Wu et al., 2023) enable inexpensive multi-agent conversation abstraction, thereby making the multi-LLMs role-playing seamless and more appealing for researchers. By having multiple models communicate in role-playing, each can more readily assimilate and replicate the persona of their assigned subject.

[5]Involves assessing an LLM's responses, behavior, and capabilities through adversarial prompting, often known as "jail-breaking"

5. Ethical considerations and Conclusion

This position paper advocates for a more nuanced perspective on the issue of cyberbullying (CB), and acknowledges the complexities beyond binary classification. Sections 2 and 3 underscore the intricacies of CB manifestations, emphasising the significance of considering diverse manifestations of CB behaviour and different participant roles when crafting evaluation cases for CB detection systems.

Moreover, to circumvent the persistent issue of CB data as studied by Emmery et al. (2021), Section 4 discusses the recent advancements in "generative" techniques as a valuable alternative. These innovative techniques empower researchers to create synthetic CB scenarios, potentially alleviating some of the daunting challenges associated with CB-scenario role-playing methods with young people.

However, it is crucial to acknowledge that while LLMs can serve as a valuable resource for generating CB scenarios, they come with inherent limitations. Leveraging multiple LLMs to simulate child participants in CB scenarios is an option worth exploring, but researchers must proceed with caution. LLMs have the capacity to produce results that lack reliability and can sometimes generate content that is biased, offensive, or factually incorrect without the oversight of human judgment. This underscores the need for robust validation mechanisms to ensure the quality and appropriateness of the generated content. Moreover, it is important to note that experiments with commercial models like Claude-2 can be ethically challenging due to the proprietary nature of the training data. Independent research should consider conducting such experiments by evaluating generated content with a diverse team of experts, including social scientists and linguists. Additionally, involving social workers from youth groups or trusted flaggers who possess a deeper understanding of interactions among young demographics can provide invaluable perspectives on the appropriateness, sensitivity, and aid in evaluating realism of the LLM-generated content in the CB context. By fostering interdisciplinary collaboration and incorporating diverse perspectives, researchers can navigate the ethical complexities associated with LLM-based CB scenario generation more effectively, ensuring the validity, reliability, and ethical integrity of their studies

Additional approaches like active learning (Settles, 2009; Mosqueira-Rey et al., 2023) in human-in-the-loop machine learning, where an entity = human (often a domain expert) is queried by the learner in control of the data (instruct-tuned LLMs) to annotate unlabeled examples (LLM generated text). Newer techniques like human-in-the-loop chain-of-thought prompting (Cai et al., 2023) too could be explored so as to help mitigate the biases or otherwise factually incorrect generated text. It is however important to note that such strategies should involve human participants who understand the domain of cyberbullying. This would generally involve humans who are (a) able to recognise such behaviour either through their past experience or witnessing such behaviour or (b) experts in the domain who study such behaviour. Moreover, instead of directly involving pre-adolescents or teenagers, such strategies should involve social youth group workers or student ambassadors or teachers, who work closely with pre-adolescents or teenagers. As they can offer nuanced insights into the needs, preferences, and sensitivities of young people. Thereby ensuring that any interventions or research initiatives are tailored to best serve their interests and well-being.

In essence, this position paper puts forth a call to action for a more comprehensive and nuanced approach to CB research. It underscores the importance of understanding the multifaceted nature of CB and encourages researchers to address the ethical and reliability challenges that come into play when utilising LLMs for data generation in this context. Moreover this paper can inform the development of a socio-linguistic framework with emphasis on (a) how and why people engage in CB behaviour, (b) what effects it has, (c) the evolving nature of CB in discourse, and (d) aligning with existing user-generated content linguistic frameworks like (Sanguinetti et al., 2020), especially in the context of vulnerable population - teenagers. Future cyberbullying research should prioritise the multidimensional nature of CB, encompassing the diverse range of participation roles and CB types to better understand this complex issue.

6. Acknowledgments

We thank the authors (Sprugnoli et al., 2018) for making their data repository[6] publicly available.

This research is supported by the Irish Research Council (IRC) Enterprise Partnership Scheme (EPS) with Google for Online Content Safety under grant number EPSPG/2021/161 and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology grant number 13/RC/2106_P2.

In addition, Kolawole Adebayo is supported by Enterprise Ireland's CareerFit-Plus Co-fund and the European Union's Horizon 2020 research and innovation programme Marie Skłodowska-Curie Grant No. 847402. Megan Reynolds and Joachim Wagner acknowledge support by the Disruptive Technologies Innovation Fund (DTIF) under the project "Cilter: Protecting Children Online" Grant No. DT 2021 0362 from the Department of Enterprise, Trade and Employment in Ireland and administered by Enterprise Ireland (EI).

[6] https://github.com/dhfbk/WhatsApp-Dataset/tree/master

7. Bibliographical References

Aggarwal, A., Maurya, K., and Chaudhary, A. (2020). Comparative study for predicting the severity of cyberbullying across multiple social media platforms. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 871–877. IEEE.

Aggarwal, P. and Zesch, T. (2022). Analyzing the real vulnerability of hate speech detection systems against targeted intentional noise. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 230–242.

Agrawal, S. and Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer.

Ahmed, T., Ivan, S., Kabir, M., Mahmud, H., and Hasan, K. (2022). Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying. *Social Network Analysis and Mining*, 12(1):99.

Ahmed, T., Kabir, M., Ivan, S., Mahmud, H., and Hasan, K. (2021). Am i being bullied on social media? an ensemble approach to categorize cyberbullying. In *2021 IEEE international conference on big data (Big data)*, pages 2442–2453. IEEE.

Alotaibi, M., Alotaibi, B., and Razaque, A. (2021). A multichannel deep learning framework for cyberbullying detection on social media. *Electronics*, 10(21):2664.

Anthropic (2023). Model card and evaluations for claude models.

Baek, J. and Bullock, L. M. (2014). Cyberbullying: a cross-cultural perspective. *Emotional and behavioural difficulties*, 19(2):226–238.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback.

Banko, M., MacKeen, B., and Ray, L. (2020). A unified taxonomy of harmful content. In *Proceedings of the fourth workshop on online abuse and harms*, pages 125–137.

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., and De Bourdeaudhuij, I. (2015). 'can i afford to help?'how affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. *Behaviour & Information Technology*, 34(4):425–435.

Bauman, S. (2015). Types of cyberbullying. *Cyberbullying: What counselors need to know*, pages 53–58.

Behzadi, M., Harris, I. G., and Derakhshan, A. (2021). Rapid cyber-bullying detection method using compact bert models. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 199–202. IEEE.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bhargava, A., Witkowski, C., Shah, M., and Thomson, M. (2023). What's the magic word? a control theory of llm prompting. *arXiv preprint arXiv:2310.04444*.

Bharti, S., Yadav, A. K., Kumar, M., and Yadav, D. (2021). Cyberbullying detection from tweets using deep learning. *Kybernetes*, 51(9):2695–2711.

Bonanno, R. A. and Hymel, S. (2013). Cyber bullying and internalizing difficulties: Above and beyond the impact of traditional forms of bullying. *Journal of youth and adolescence*, 42:685–697.

Bretschneider, U. and Peters, R. (2016). Detecting cyberbullying in online communities. *Proceedings of the 24th European Conference on Information Systems (ECIS 2016), Research Papers*. Paper 61.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Buglass, S. L., Abell, L., Betts, L. R., Hill, R., and Saunders, J. (2021). Banter versus bullying: A university student perspective. *International journal of bullying prevention*, 3:287–299.

Cai, Z., Chang, B., and Han, W. (2023). Human-in-the-loop through chain-of-thought. *arXiv preprint arXiv:2306.07932*.

Cao, R. and Lee, R. K.-W. (2020). Hategan: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338.

Cappadocia, M. C., Craig, W. M., and Pepler, D. (2013). Cyberbullying: Prevalence, stability, and risk factors during adolescence. *Canadian Journal of School Psychology*, 28(2):171–192.

Cassidy, W., Faucher, C., and Jackson, M. (2013). Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice. *School psychology international*, 34(6):575–612.

Chan, T. K., Cheung, C. M., Benbasat, I., Xiao, B., and Lee, Z. W. (2022). Bystanders join in cyberbullying on social networking sites: the deindividuation and moral disengagement perspectives. *Information Systems Research*.

Chan, T. K., Cheung, C. M., and Lee, Z. W. (2021). Cyberbullying on social networking sites: A literature review and future research directions. *Information & Management*, 58(2):103411.

Chan, T. K., Cheung, C. M., and Wong, R. Y. (2019). Cyberbullying on social networking sites: The crime opportunity and affordance perspectives. *Journal of Management Information Systems*, 36(2):574–609.

Charisi, V., Chaudron, S., Di Gioia, R., Vuorikari, R., Escobar Planas, M., Sanchez, M. J. I., and Gomez Gutierrez, E. (2022). Artificial intelligence and the rights of the child: Towards an integrated agenda for research and policy. Technical report, Joint Research Centre (Seville site).

Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E. D., Stringhini, G., Vakali, A., and Kourtellis, N. (2019). Detecting cyberbullying and cyberaggression in social media. *ACM Transactions on the Web (TWEB)*, 13(3):1–51.

Dadvar, M. and Eckert, K. (2020). Cyberbullying detection in social networks using deep learning based models. In *Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22*, pages 245–255. Springer.

Davidson, J., Livingstone, S., Jenkins, S., Gekoski, A., Choak, C., Ike, T., and Phillips, K. (2019). Adult online hate, harassment and abuse: a rapid evidence assessment. Technical report, London School of Economics (LSE).

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the international AAAI conference on web and social media*, 11(1):512–515.

Elsafoury, F., Katsigiannis, S., Wilson, S. R., and Ramzan, N. (2021). Does bert pay attention to cyberbullying? In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1900–1904.

Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., Desmet, B., Hoste, V., and Daelemans, W. (2021). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. *Language Resources and Evaluation*, 55:597–633.

Espelage, D. L. and Swearer, S. M. (2003). Research on school bullying and victimization: What have we learned and where do we go from here? *School psychology review*, 32(3):365–383.

Ey, L.-A., Taddeo, C., and Spears, B. (2015). Cyberbullying and primary-school aged children: The psychological literature and the challenge for sociology. *Societies*, 5(2):492–514.

Fang, Y., Yang, S., Zhao, B., and Huang, C. (2021). Cyberbullying detection in social networks using bi-gru with self-attention mechanism. *Information*, 12(4):171.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Fredstrom, B. K., Adams, R. E., and Gilman, R. (2011). Electronic and school-based victimization: Unique contexts for adjustment difficulties during adolescence. *Journal of youth and adolescence*, 40:405–415.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12.

Inc., P. J. F. (2020). Perverted-justice.com archives. `http://www.perverted-justice.com/?archive=byUserVotes`.

Jain, V., Kumar, V., Pal, V., and Vishwakarma, D. K. (2021). Detection of cyberbullying on social media using machine learning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1091–1096. IEEE.

Jicol, C., Feltham, J., Yoon, J., Proulx, M. J., O'Neill, E., and Lutteroth, C. (2022). Designing and assessing a virtual reality simulation to build resilience to street harassment. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Kaggle (2018). Tweets dataset for detection of cyber-trolls.

Kane, G. C., Alavi, M., Labianca, G., and Borgatti, S. P. (2014). What's different about social media networks? a framework and research agenda. *MIS quarterly*, 38(1):275–304.

Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T., and Hale, S. A. (2022). Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate.

Kowalski, R. (2018). Cyberbullying. In *The Routledge international handbook of human aggression*, pages 131–142. Routledge.

Leung, A. N., Wong, N., and Farver, J. M. (2018). You are what you read: The belief systems of cyber-bystanders on social networking sites. *Frontiers in psychology*, 9:365.

Li, H., Guo, D., Fan, W., Xu, M., Huang, J., Meng, F., and Song, Y. (2023). Multi-step jailbreaking privacy attacks on chatgpt.

Lobe, B., Velicu, A., Staksrud, E., Chaudron, S., and Di Gioia, R. (2021). How children (10-18) experienced online risks during the covid-19 lockdown - spring 2020. *JRC Publications Repository*, JRC(KJ-NA-30584-EN-N (online),KJ-NA-30584-EN-C (print)).

Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., and Choo, K.-K. R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. *Concurrency and Computation: Practice and Experience*, 32(23):e5627.

McFarland, L. A. and Ployhart, R. E. (2015). Social media: A contextual framework to guide research and practice. *Journal of applied psychology*, 100(6):1653.

Menesini, E., Nocentini, A., Palladino, B. E., Scheithauer, H., Schultze-Krumbholz, A., Frisén, A., Sofia, B., Luik, P., Naruskov, K., Ortega, R., et al. (2013). Definitions of cyberbullying. In *Cyberbullying through the new media*, pages 23–36. Psychology Press.

Menini, S., Moretti, G., Corazza, M., Cabrio, E., Tonelli, S., and Villata, S. (2019). A system to monitor cyberbullying based on message classification and social network analysis. In *Proceedings of the third workshop on abusive language online*, pages 105–110.

Meta (2023). Community standards enforcement. https://transparency.fb.com/reports/community-standards-enforcement/bullying-and-harassment/facebook/.

Milosevic, T., Laffan, D., and O'Higgins Norman, J. (2021). Kidicoti: Kids' digital lives in covid-19 times: A study on digital practices, safety and wellbeing. key findings from ireland. https://antibullyingcentre.ie/wp-content/uploads/2021/12/Short-report_Covid_for-media_TM_with-Author-names-1-2.pdf.

Monks, C. P., Robinson, S., and Worlidge, P. (2012). The emergence of cyberbullying: A survey of primary school pupils' perceptions and experiences. *School Psychology International*, 33(5):477–491.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Muneer, A. and Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, 12(11):187.

Nadali, S., Murad, M. A. A., Sharef, N. M., Mustapha, A., and Shojaee, S. (2013). A review of cyberbullying detection: An overview. In *2013 13th International Conference on Intellient Systems Design and Applications*, pages 325–330. IEEE.

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., and Scheidt, P. (2001). Bullying behaviors among us youth: Prevalence and association with psychosocial adjustment. *Jama*, 285(16):2094–2100.

O., B., A., C., A., K., Flynn, E., , and Messena (2023). Keeping kids safer online. online safety matters. trends and usage report academic year 2022/2023. https://www.cybersafekids.ie/wp-content/uploads/2023/09/CSK_Data-Trends-Report-2023-V2-Web-Version.pdf.

Ollagnier, A., Cabrio, E., Villata, S., and Blaya, C. (2022). CyberAgressionAdo-v1: a dataset of annotated online aggressions in French collected through a role-playing game. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 867–875, Marseille, France. European Language Resources Association.

Olweus, D. (2013). Victimization by peers: Antecedents and long-term outcomes. *Social withdrawal, inhibition, and shyness in childhood*, pages 315–341.

Opitz, J. and Burst, S. (2019). Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Paruchuri, V. L. and Rajesh, P. (2022). Cybernet: a hybrid deep cnn with n-gram feature selection for cyberbullying detection in online social networks. *Evolutionary Intelligence*, pages 1–15.

Patchin, J. W. and Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth violence and juvenile justice*, 4(2):148–169.

Paul, S. and Saha, S. (2022). Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification. *Multimedia Systems*, 28(6):1897–1904.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., Showk, S. E., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations.

Perez, F. and Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models.

Pericherla, S. and Ilavarasan, E. (2021). Transformer network-based word embeddings approach for autonomous cyberbullying detection. *International Journal of Intelligent Unmanned Systems*.

Phanomtip, A., Sueb-in, T., and Vittayakorn, S. (2021). Cyberbullying detection on tweets. In *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 295–298. IEEE.

Pradhan, A., Yatam, V. M., and Bera, P. (2020). Self-attention for cyberbullying detection. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pages 1–6. IEEE.

Pyżzalski, J. (2013). From cyberbullying to electronic aggression: Typology of the phenomenon. In *Emotional and Behavioural Difficulties Associated with Bullying and Cyberbullying*, pages 79–91. Routledge.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Raj, C., Agarwal, A., Bharathy, G., Narayan, B., and Prasad, M. (2021). Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques. *Electronics*, 10(22):2810.

Randall, P. (2003). *Adult bullying: Perpetrators and victims*. Routledge.

Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of nlp models with checklist.

Rizos, G., Hemker, K., and Schuller, B. (2019). Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 991–1000.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., and Pierrehumbert, J. (2021). Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., and Zeldes, A. (2020). Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5240–5250, Marseille, France. European Language Resources Association.

Settles, B. (2009). Active learning literature survey. Technical report, University of Wisconsin-Madison.

Slonje, R., Smith, P. K., and Frisén, A. (2013). The nature of cyberbullying, and strategies for prevention. *Computers in human behavior*, 29(1):26–32.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., and Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385.

Smith, P. K. and Steffgen, G. (2013). *Cyberbullying through the new media: Findings from an international network*. Psychology Press.

Song, J. and Oh, I. (2018). Factors influencing bystanders' behavioral reactions in cyberbullying situations. *Computers in Human Behavior*, 78:273–282.

Sprugnoli, R., Menini, S., Tonelli, S., Oncini, F., and Piras, E. (2018). Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59.

Steer, O. L., Betts, L. R., Baguley, T., and Binder, J. F. (2020). "i feel like everyone does it"-adolescents' perceptions and awareness of the association between humour, banter, and cyberbullying. *Computers in Human Behavior*, 108:106297.

Sticca, F., Ruggieri, S., Alsaker, F., and Perren, S. (2013). Longitudinal risk factors for cyberbullying in adolescence. *Journal of community & applied social psychology*, 23(1):52–67.

Thomas, H. J., Connor, J. P., and Scott, J. G. (2015). Integrating traditional bullying and cyberbullying: challenges of definition and measurement in adolescents–a review. *Educational psychology review*, 27:135–152.

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior*, 26(3):277–287.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Tripathy, J. K., Chakkaravarthy, S. S., Satapathy, S. C., Sahoo, M., and Vaidehi, V. (2022). Albert-based fine-tuning model for cyberbullying analysis. *Multimedia Systems*, 28(6):1941–1949.

Urbaniak, R., Ptaszyński, M., Tempska, P., Leliwa, G., Brochocki, M., and Wroczyński, M. (2022). Personal attacks decrease user activity in social networking platforms. *Computers in Human Behavior*, 126:106972.

Vaillancourt, T., McDougall, P., Hymel, S., Krygsman, A., Miller, J., Stiver, K., and Davis, C. (2008). Bullying: Are researchers and children/youth talking about the same thing? *International Journal of Behavioral Development*, 32(6):486–495.

Van Bruwaene, D., Huang, Q., and Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, 54:851–874.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.

Verma, K., Popović, M., Poulis, A., Cherkasova, Y., Mazzone, A., Milosevic, T., Davis, B., et al. (2022). Leveraging machine translation for cross-lingual fine-grained cyberbullying classification amongst pre-adolescents. *Natural Language Engineering*, pages 1–23.

Wade, A. and Beran, T. (2011). Cyberbullying: The new era of bullying. *Canadian Journal of School Psychology*, 26(1):44–61.

Wang, J., Fu, K., and Lu, C.-T. (2020). Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1699–1708. IEEE.

Wang, P., Guo, Z., Huang, L., and Cho, J.-H. (2023). Authentic dialogue generation to improve youth's awareness of cybergrooming for online safety. TechRxiv, Authorea Preprints.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Wei, A., Haghtalab, N., and Steinhardt, J. (2023). Jailbroken: How does llm safety training fail?

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Workshop, B., :, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., Toni, F. D., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Werra, L. V., Weber, L., Phan, L., allal, L. B., Tanguy, L., Dey, M., Muñoz, M. R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M. T.-J., Vu, M. C., Jauhar, M. A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R. L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S. H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T. T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D. E., Salesky, E., Mielke, S. J., Lee, W. Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J. A., Rozen, J., Gao, L., Sutawika, L., Bari, M. S., Al-shaibani, M. S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S. H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H. W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P. F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G. I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J. Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C. M., McDuff, D., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D. A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J. B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A. R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D. L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrimann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H. U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J. D., Sivaraman, K. R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M. H., Takeuchi, M., Pàmies, M., Castillo, M. A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., Wolf, M. D., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N. M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S. S., Mishra, S., Kiblawi, S., Ott, S., Sang-aroonsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y. S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., and Wolf, T. (2023). Bloom: A 176b-parameter open-access multilingual language model.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. (2023). Autogen: Enabling next-gen llm applications via multi-agent conversation.

Wu, T., Ribeiro, M. T., Heer, J., and Weld, D. S. (2019). Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Yadav, J., Kumar, D., and Chauhan, D. (2020). Cyberbullying detection using pre-trained bert model. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1096–1100. IEEE.

Zhang, H., Song, H., Li, S., Zhou, M., and Song, D. (2022). A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*.

Zhao, Z., Gao, M., Luo, F., Zhang, Y., and Xiong, Q. (2020). Lshwe: improving similarity-based word embedding with locality sensitive hashing for cyberbullying detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Zhu, Z., Xue, Y., Chen, X., Zhou, D., Tang, J., Schuurmans, D., and Dai, H. (2023). Large language models can learn rules. *arXiv preprint arXiv:2310.07064*.

Ziems, C., Vigfusson, Y., and Morstatter, F. (2020). Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 808–819.

A. Tables

Study	Dataset	Dataset Type	Harm Type
(Phanomtip et al., 2021)	(Waseem and Hovy, 2016; Wulczyn et al., 2017; Reynolds et al., 2011)	Binary	Personal Attacks; Binary Bullying; Hate-speech
(Zhao et al., 2020)			
(Aggarwal et al., 2020)			
(Paruchuri and Rajesh, 2022)			
(Yadav et al., 2020)	(Agrawal and Awekar, 2018; Davidson et al., 2017; Waseem and Hovy, 2016; Chatzakou et al., 2019; Bretschneider and Peters, 2016)	Binary	Hate-speech; Aggression; Offensive language
(Pradhan et al., 2020)	(Wulczyn et al., 2017; Reynolds et al., 2011; Waseem and Hovy, 2016)	Binary	Hate-speech and Personal attacks
(Fang et al., 2021)	(Reynolds et al., 2011; Waseem and Hovy, 2016)	Binary	Hate-Speech; Binary Bullying
(Tripathy et al., 2022)	(Davidson et al., 2017)	Binary	Hate-Speech
(Muneer and Fati, 2020) (Behzadi et al., 2021)	(Founta et al., 2018)	Binary	Hate-Speech
(Paul and Saha, 2022)	(Waseem and Hovy, 2016; Davidson et al., 2017)	Binary	Hate-Speech; Offensive language
(Agrawal and Awekar, 2018)			
(Wang et al., 2020)			
(Elsafoury et al., 2021)			
(Bharti et al., 2021)			
(Dadvar and Eckert, 2020) (Raj et al., 2021)	(Waseem and Hovy, 2016)	Binary	Hate-speech
(Alotaibi et al., 2021)			
(Menini et al., 2019)			
(Lu et al., 2020)	(Waseem and Hovy, 2016; Wulczyn et al., 2017; Kaggle, 2018)	Binary	Hate-speech; Offensive language
(Jain et al., 2021)	(Waseem and Hovy, 2016; Wang et al., 2020)	Binary	Hate-Speech; Offensive language
(Ahmed et al., 2021)	(Wulczyn et al., 2017)	Binary	Personal attacks
(Pericherla and Ilavarasan, 2021)	(Wulczyn et al., 2017)	Binary	Personal attacks
(Chatzakou et al., 2019)	Custom (No Annotation Guideline)	Binary	Hate-speech & Abusive language
(Ahmed et al., 2022)	(Wang et al., 2020; Waseem and Hovy, 2016)	Binary	Hate-speech & Offensive language

Table 1: List of studies that claim to detect “cyberbullying” but use out-of-domain datasets

Pompt-1: You are presented with conversation between school-age children. They are role-playing conversation being given this scenario - <Scenario> Your shy male classmate has a great passion for classical dance. Usually he does not talk much, but today he has decided to invite the classroom to watch him for his ballet show </scenario> See a sample of their conversation <Conversation-Chain> ... </Conversation-Chain> What do you learn from this conversation?
Claude-2: I learned a few things from this conversation: (1) Person 1 is brave for pursuing his passion for ballet even though it goes against gender stereotypes. He invites his classmates to his recital despite being shy and embarrassed. (2) Persons 3 and 7 react with prejudice, making fun of Person 1 for liking ballet and calling him names. Their comments are mean-spirited and close-minded. (3) Persons 2 and 8 stand up for Person 1. They see nothing wrong with him liking ballet and encourage him to keep dancing. (4) Pursuing interests outside of gender norms can lead to bullying, but having even a couple supportive friends makes a big difference. We should accept people's diverse interests without judgement. (5) Persons 3 and 7 try to shame Person 1, but he remains steadfast. Standing up to bullying requires courage. (6) Overall, this shows how biases against gender non-conformity can lead to cruel teasing. But supportive friends and self-confidence help overcome negativity. We should be more open-minded and not judge people for their hobbies or interests
Llama2Chat$_{\mathbf{7B}}$: This conversation is not appropriate or respectful. It contains harmful and toxic language, including homophobic slurs and gendered insults. The use of such language can create a hostile environment and contribute to a culture of discrimination and marginalization. Additionally, the conversation implies that certain activities or interests are inherently "feminine" or "masculine," which reinforces harmful gender stereotypes. It is important to recognize that everyone should be treated with dignity and respect, regardless of their gender expression, sexual orientation, or interests. Everyone deserves the opportunity to pursue their passions without fear of ridicule or harassment. Instead of engaging in hurtful conversations, it's important to promote inclusivity, diversity, and respect for all individuals. We can encourage open-mindedness and empathy towards others, and work towards creating a society where everyone can feel safe and valued.

Table 2: Response to Prompt-1 by Claude-2 and Llama2Chat.

Pompt-2: Have you detected any instances of cyberbullying in this conversation?
Claude-2: Yes, there are a few instances of cyberbullying evident in this conversation: (1) Person 3 threatens to print t-shirts with embarrassing images of Person 1 in a tutu. Creating and distributing humiliating material online could be considered cyberbullying. (2) Persons 3 and 7 use mean and derogatory language like 'sissy', 'fag', and comments about Person 1 playing with females. The harassing comments attempt to shame Person 1 online. (3) Persons 3 and 7 say they will attend the recital only to laugh at and mock Person 1. Coordinating to embarrass someone in real life based on online activities could cross over into cyberbullying. (4) Person 7 calls Person 1 an 'idiot' who will 'make a fool of himself'. Insulting language used online can constitute cyberbullying. (5) The overall tone of repeated insults, name-calling and attempts to shame Person 1 for his interests create a cyberbullying dynamic even though the children know each other in real life. (6) So while this isn't a clear cut case, there are definite elements of cyberbullying present in the ways Persons 3 and 7 try to shame, embarrass and put down Person 1 online. The harassment attempts to damage his reputation and make him feel bad about his interests.
Llama2Chat$_{\mathbf{7B}}$: Yes, I did detect some instances of cyberbullying in the conversation you provided. Specifically, the following statements could be considered cyberbullying: "You're so effeminate that there are more pansies around you than in a garden."; "I'll have T-shirts printed with you wearing a tutu."; "The only fag who will dance with the girls."; "These statements contain derogatory language and stereotypes that are intended to belittle and shame the person being addressed. They also perpetuate harmful and negative attitudes towards people who identify as LGBTQ+." Cyberbullying can have serious consequences for the mental health and well-being of those who are targeted. It's important to be aware of these behaviors and to take steps to prevent them in our online interactions.

Table 3: Response to Prompt-2 by Claude-2 and Llama2Chat.

Pompt-3: What categorization can you give the various instances of bullying?
Claude-2: Based on my analysis, I would categorize the instances of bullying in this conversation as: • Verbal bullying: Name-calling, insults, and derogatory language used by Persons 3 and 7, like 'sissy', 'fag', calling Person 1 an 'idiot', etc • Relational bullying: Threats to socially exclude or embarrass Person 1. For example, Persons 3 and 7 say they will attend the recital only to mock and laugh at him. • Cyberbullying: Person 3's threat to print embarrassing t-shirts to distribute, which would humiliate Person 1 online/in real life. Also the overall harassment in the chat space. • Prejudiced bullying:Targeting and insulting Person 1 based on gender stereotypes and prejudices against males doing ballet. • Public humiliation: Persons 3 and 7 coordinate plans to publicly embarrass Person 1 by laughing at him and calling names during his recital. • Criticism of hobbies/interests: Making Person 1 feel ashamed or attacked for his passion for ballet • Retaliation: The bullying seems partially intended to get back at Person 1 for challenging gender norms with his interest in ballet. So in summary, this bullying employs verbal attacks, relational exclusion, cyber harassment, prejudice, public embarrassment, and criticism of Person 1's interests - all intended to shame him for his hobby.
Llama2Chat$_{7B}$: In this conversation, there are several instances of bullying that can be categorized in different ways. Here are some examples: • Verbal bullying: This involves using words to hurt, intimidate, or threaten someone. Examples of verbal bullying in this conversation include: * Calling someone a "sissy" or "effeminate" * Using derogatory language or slurs * Making fun of someone's appearance or behavior • Social bullying: This involves excluding someone from a group or spreading rumors to isolate them. Examples of social bullying in this conversation include: * Saying that someone is "not one of us" * Implying that someone is weird or doesn't fit in * Encouraging others to exclude someone from activities or conversations • Physical bullying: This involves physical violence or threats of violence. While there are no explicit examples of physical bullying in this conversation, some comments could potentially escalate into physical altercations. For example: * Threatening to "print T-shirts" with someone's picture wearing

Table 4: Response to Prompt-3 by Claude-2 and Llama2Chat.

Pompt-4: Create a similar conversation
Claude-2: Here is a similar constructed conversation with the same bullying dynamics: • Person 1: Hi everyone. I wanted to invite you all to the craft fair this weekend. I'll be selling handmade jewelry and knitted scarves that I made myself. I'm a bit nervous to put myself out there, but I've been working hard on this for months. • Person 3: Jewelry and knitting? Isn't that stuff for old ladies? What are you, a grandma? • Person 7: Haha Person 1, what loser hobbies. I can't believe you spend time knitting scarves, how embarrassing! • Person 2: Don't listen to them Person 1! I think it's really cool you make your own jewelry and knit. I'll definitely stop by your booth at the craft fair. • Person 8: Seriously you guys, stop being jerks. So what if Person 1 likes crafts? They're fun hobbies and I'm sure his stuff will be awesome. I can't wait to see it! • Person 3: I'll make sure to bring my friends to point and laugh at Person 1's lame grandma booth. What a joke! • Person 7: Hahaha yes! We can take pictures and post them too to show everyone what a weirdo Person 1 is for knitting. • Person 1: There's nothing wrong with enjoying crafts and handmade items. I don't see why you need to be so mean about my hobbies. • Person 2: Don't pay any attention to these immature bullies Person 1. Keep doing what you love. I'll help promote your booth! • Person 8: Exactly, don't let them get you down. Your interests are cool and unique. Forget the haters!-
Llama2Chat$_{7B}$: Sure, here's a conversation between a group of students discussing a new student who has just transferred to their school: • Student 1: "Have you guys seen the new kid yet? He's totally awkward and doesn't know anyone." • Student 2: "Yeah, I saw him sitting alone at lunch yesterday. He looked really sad." • Student 3: "I heard he's from a different country, so maybe that's why he's struggling to fit in." • Student 4: "Whatever, dude. He's still super weird. I bet he's never had a real friend before." • Student 5: "Guys, come on. Let's not be mean. We don't know anything about him yet. Maybe he's really cool once you get to know him." • Student 6: "Yeah, and even if he is a little different, that doesn't give us the right to bully him. Remember how we all felt when we first started here?" • Student 1: "Okay, okay. You're right. Let's try to be nice to him"

Table 5: Response to Prompt-4 by Claude-2 and Llama2Chat.

B. Scenario

The scenario presented below has been devised by (Sprugnoli et al., 2018)

```
  Your shy male classmate has a
great passion for classical dance.
Usually he does not talk much, but
today he has decided to invite
the classroom to watch him for his
ballet show.
```

C. Conversation-Chain

The conversation provided in (Sprugnoli et al., 2018) is originally in the Italian language and has been translated into English by (Verma et al., 2022).

```
{
    Person1 : ``Hi . . . I wanted to invite you to the ballet recital on Saturday, March 3 . . . it would make me very happy . . . I am a little embarrassed'',
    Person2: ``I will ask my mom and I will let you know'',
    Person1 : ``Thanks Person2!'',
    Person3 : ``Ah a dance recital'',
    "Person3" : "You who are a male",
    "Person7" : "Actually, does it seem normal to you?",
    "Person8" : "Thanks Person1 I will definitely be there",
    "Person3" : "No, it is for girls",
    "Person3" : "Then I imagine him wearing leggings",
    "Person3" : "What about the tutu?",
    "Person7" : "You, who dances like an idiot ",
    "Person1" : "What's wrong with that ?",
    "Person3" : "You on tips (en pointe ?)",
    "Person8" : "Person1, don't listen to them.",
    ``Person3'' : ``Go play soccer you sissy'',
    "Person7" : "Person3  is right ! ! ! ! !",
    "Person3" : "Of course",
    "Person3" : "! ! ! ! !",
    "Person3" : "But then with whom, with other females",
    "Person3" : "females like you by the way",
    "Person7": "Oooooooooooooo \emoji{grinning-face}",
    "Person8" : "Leave them alone, if you like to dance, dance, if you like to sing, sing, free of prejudice",
    "Person3" : "Yes, and you go play rugby.",
    "Person3" : "Same thing",
    "Person7" : "But are you sure you have been selected for the recital or do you infiltrate to make an impression",
    "Person7" : "You have been brilliantly shamed",
    "Person1" : "Thank you. But if I like to dance and Person3 likes to play rugby, what would be the problem?",
    "Person3" : "That ballet is not for boys",
    "Person3" : "Of course",
    "Person8" : "And people who NEVER, EVER, EVER mind their own business.",
    "Person3" : "You're so effeminate that there are more pansies around you than in a garden.",
    "Person2" : "But why do you care if he dances ? If you don't like dancing don't even go to see him . I'll be there for sure",
    "Person3" : "Me too.",
    "Person3" : "Just to laugh and make fun of him.",
    "Person8" : "You are a MONSTER",
    "Person2" : "It's nonsense, leave him alone",
    "Person8" : "Exactly",
    "Person7" : "Of course I'll be there when Person1 will make a fool of himself.",
    "Person8" : "Come on Person1, I believe in you",
    "Person7" : "The only fag who will dance with the girls",
    "Person2" : "But I'd like to see you dance! I'm sure he's better than you.",
    "Person3" : "I'll have T-shirts printed with you wearing a tutu .",
    "Person7" : "Great Person3  \emoji{grinning-face}",
    "Person8" : "he is very lucky instead",
    "Person3" : "To have us insult him?",
    "Person3" : "I would say yes"
}
```