AEGIS

Current Criminology

2026 / 1 (4) / 19-36

RESEARCH ARTICLE

Anti-Cyberbullying AI Chatbots: A Systematic Review of Criminological Theories, Evidence, and Criminal Justice Applications

Zariatu IBRAHIM[1]; Hua ZHONG[1*]

[1] Department of Sociology, The Chinese University of Hong Kong

Corresponding author: Hua ZHONG (Sarazhong@cuhk.edu.hk)

Author email: Zariatu IBRAHIM (Zibrahim@link.cuhk.edu.hk); Hua ZHONG (Sarazhong@cuhk.edu.hk)

ORCID: Zariatu IBRAHIM 0009-0008-4789-8184

Abstract: This systematic review synthesizes evidence on AI chatbots for cyberbullying prevention and intervention from a criminological perspective, with explicit attention to Asia Pacific applicability. Following PRISMA 2020 guidelines, we identified 25 studies (2018-2025) across seven databases and appraised each using criteria adapted from MMAT and JBI tools, assigning every source an evidentiary weight (High, Moderate, Low, or Indicative proof-of-concept). Four chatbot roles emerged: real-time detection, victim support, preventive education, and bystander intervention training. Disaggregated analysis tempers headline performance claims. The frequently cited 89-99% accuracy range reflects laboratory classification on benchmark datasets, predominantly English-language Western corpora, evaluated under balanced class distributions and within-dataset conditions. No included study reported AUC or class-imbalance-adjusted metrics, employed temporal validation, or evaluated systems in real-world deployment with adolescents. Where precision and recall were disaggregated, asymmetric error profiles emerged with criminologically distinct social costs that single-metric reporting obscures. Only three studies qualified as High evidentiary weight, all addressing user-centred design in Western individualist contexts; victim support claims rest predominantly on Indicative proof-of-concept demonstrations. Geographic and linguistic concentration is a defining characteristic of the evidence base: zero studies have validated systems for Southeast Asian languages, no participatory design has engaged Asian youth, and no work has evaluated Asia Pacific platforms (WeChat, LINE, KakaoTalk). Drawing on Asian and Southern criminology, we argue that the uncritical transfer of Western-validated systems risks ineffectiveness, discriminatory false positives, and failure to detect culture-specific harms. The paper offers five criminological practice implications, each framed as theoretically informed extrapolations requiring local validation. AI chatbots are positioned as potentially valuable complementary tools within human-led prevention ecosystems, conditional on rigorous local research, transparent metric reporting, and sustained interdisciplinary collaboration.

Keywords: AI chatbot, anti-cyberbullying, systematic review, Asian criminology, Southern criminology, Asia Pacific

Declaration of Interests: The authors report no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the authors used DeepSeek to revise and improve the wording of selected sections of the manuscript. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Introduction

The proliferation of digital technologies and social media platforms has irrevocably transformed the social landscape of adolescence across the Asia Pacific region, creating unprecedented opportunities for connection, learning, and self-expression. However, this hyper-connected environment has also given rise to a pervasive and insidious form of criminal victimization: cyberbullying. Characterised by the intentional and repeated infliction of harm through electronic means, cyberbullying has escalated into a critical global public safety and criminal justice issue, with documented increases in prevalence and severity across diverse cultural contexts in Asia, including Hong Kong, Taiwan, Mainland China, South Korea, Japan, and Southeast Asian nations (Chen & Chen, 2020; Sidhu & Sidhu, 2025; Chan & Wong, 2020). The region's unique characteristics-high smartphone penetration, collectivist cultural norms that may discourage reporting, and linguistic diversity pose distinct challenges that require contextually grounded criminological solutions.

The impacts of cyberbullying extend beyond individual psychological distress to encompass broader criminological concerns: it represents a form of repeat victimization that inflicts significant harm on victims, including heightened anxiety, depression, and social isolation, while also undermining digital citizenship and normalizing online deviance (Chan & Wong, 2020; Barlett & Gentile, 2022). From a criminological perspective, cyberbullying has emerged as a persistent challenge demanding innovative crime prevention approaches that bridge digital and physical intervention spaces.

Conventional approaches to bullying prevention and intervention-encompassing school-wide policies, awareness programs, static educational materials, and reactive responses-have demonstrated significant limitations in addressing this digital crime phenomenon. As noted by Lian et al. (2023), these methods often lack the interactivity, immediacy, and personalisation required to effectively engage digital-native youth. They are typically deployed as broad, one-time interventions, failing to provide the real-time, situational crime prevention needed during a bullying episode. Furthermore, profound barriers such as social stigma, fear of retaliation, and mistrust in formal authority contribute to drastic underreporting, leaving a vast majority of incidents unaddressed and allowing harmful behaviours to normalise within peer networks (Henry et al., 2025; Hedderich et al., 2024; Kasturiratna et al., 2025). In collectivist Asian societies, where 'saving face' and avoiding family shame are paramount concerns, these reporting barriers are particularly pronounced, often silencing victims who might otherwise seek intervention.

In this complex scenario, Artificial Intelligence (AI), particularly in the form of advanced conversational agents or chatbots, has emerged as a potentially valuable crime-prevention tool. AI-driven chatbots offer promise by providing scalable, accessible, and adaptive systems capable of operating across the spectrum of cyberbullying mitigation-from primary prevention and real-time detection to immediate intervention and post-incident victim support (Spytska, 2025; Sidhu & Sidhu, 2025). These systems can provide 24/7 support, deliver personalised coping strategies based on sentiment analysis, empower bystanders through guided training, and quickly and consistently identify harmful content. However, as Liu and Travers (2018) emphasize in their foundational work on comparative criminology in Asia, criminological research and criminal justice practice in the region cannot simply import Western frameworks and technologies uncritically. The region's distinct legal systems, collectivist cultural values prioritizing social harmony and family honor, and diverse digital platform ecologies require contextualized approaches grounded in local knowledge and priorities. Similarly, Carrington et al. (2016, 2019) argue in articulating Southern criminology that theories and

interventions developed in Global North contexts may mischaracterize or overlook dynamics specific to the Global South, including much of the Asia Pacific.

However, the integration of sophisticated AI into the sensitive domain of youth victimization and digital crime is not merely a technical challenge; it is a profound socio-technical endeavor laden with ethical, practical, and theoretical complexities (David et al., 2024; Marshall et al., 2025). This literature review seeks to move beyond a superficial cataloguing of technological features to conduct a systematic, theoretically grounded synthesis of existing research. The purpose is to answer pivotal questions: What specific roles and capabilities do AI chatbots demonstrate in detecting, preventing, and intervening in cyberbullying? How does their reported effectiveness align with the mechanisms of offending and victimisation predicted by criminological theories? What are the foremost challenges and limitations-technical, ethical, and practical that constrain their current application and future potential? And most critically, what are the implications for criminological practice and criminal justice responses in the Asia Pacific region, given that the evidence base derives predominantly from Western contexts?

By framing the review within foundational criminological models and subjecting the evidence to critical in-depth analysis, this paper aims to provide a nuanced understanding not only of what AI chatbots can do but also of how and why they might succeed or fail in disrupting the intricate social dynamics of online aggression. The paper makes two distinctive contributions: first, a theoretically grounded, methodologically transparent synthesis of existing evidence from a criminological perspective; and second, a practice-focused framework that translates findings into contextualized guidance for criminal justice professionals while explicitly acknowledging the limitations of extrapolating from a predominantly Western evidence base to diverse Asia Pacific contexts.

Theoretical Frameworks: Understanding the Roots of Cyberbullying Perpetration and Victimization

This study adopts a broad definition of cyberbullying as follows: An aggressive, intentional act carried out by a group or individual through electronic means of communication (e.g., computers and cell phones) against a group or individual who cannot easily defend themselves from such victimization (Chan & Wong, 2020). Effective crime prevention and intervention require a foundational understanding of the causative and contextual factors behind cyberbullying perpetration and victimization. This study integrates insights from several key criminological and victimological perspectives.

Perpetrator-Focused Theories

Motivation and Propensity. The behavior of potential cyberbullies can be understood through complementary criminological lenses. The General Theory of Crime posits that low self-control is a primary driver, leading individuals to pursue immediate gratification through online aggression without considering the long-term consequences (Bilewicz et al., 2021; Gottfredson & Hirschi, 1990). Social Learning Theory suggests that aggressive behaviors are learned and reinforced through observation, imitation, and perceived rewards within one's social environment, including online spaces (Pimpista et al., 2020; Bandura, 1978). Furthermore, Strain Theory highlights how negative emotions arising from stress, perceived injustice, or failure to achieve goals can create pressure that is displaced into online hostility (Agnew, 1992). These theories collectively provide the 'motivation and propensity' for harmful behavior. Critically, in Asian educational contexts characterized by intense academic pressure, Strain Theory may have explanatory power, as students experiencing academic failure may displace frustration through online aggression (Chu et al., 2023).

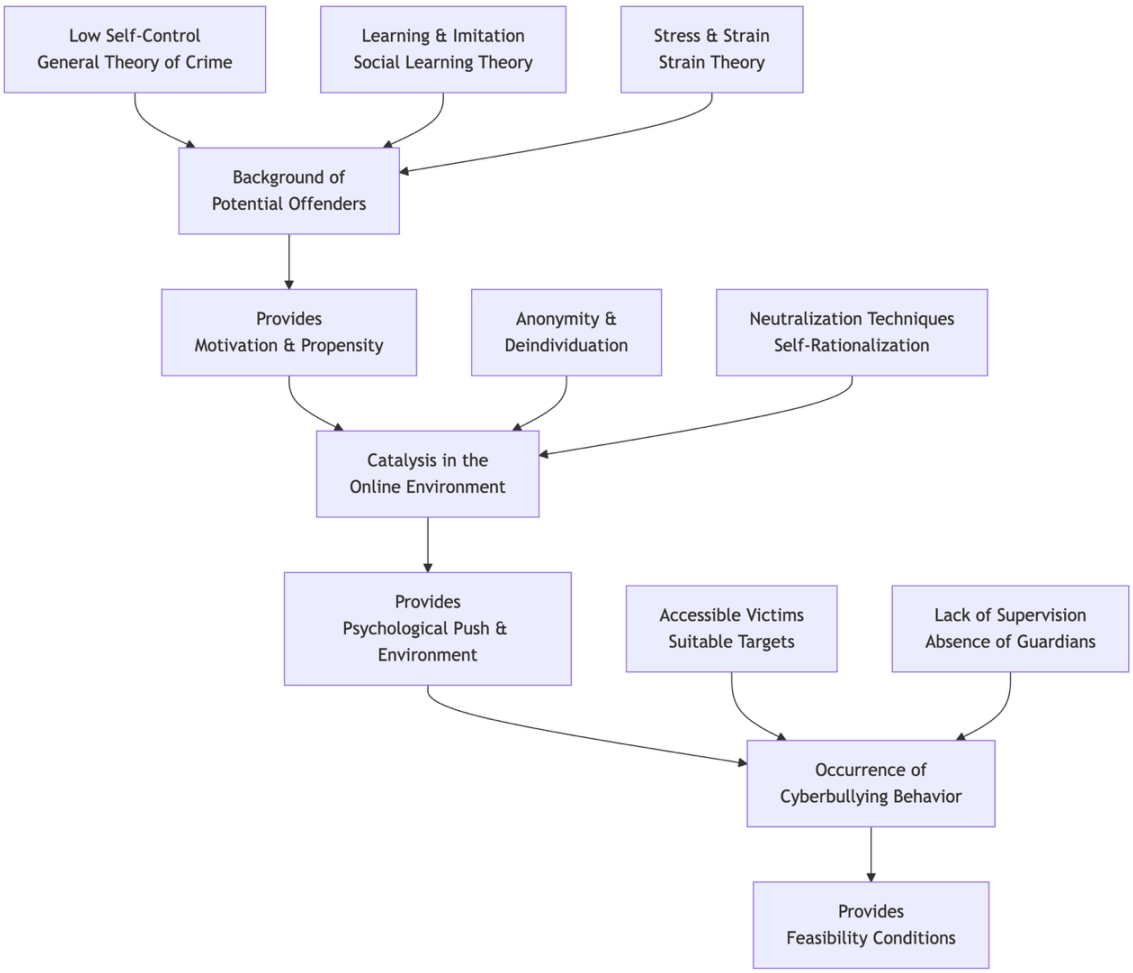
The applicability of these predominantly Western-origin theories to Asia Pacific contexts requires critical examination. As Carrington et al. (2016) argue, criminological theories developed in Anglo-American contexts may emphasize individual-level factors (e.g., self-control, personal victimization risk) while underestimating collective-level dynamics more salient in collectivist Asian societies-including family honor, interdependent self-construal, hierarchical authority

relationships, and community-level social control. The General Theory of Crime's focus on individual self-control, for example, may need to be adapted to account for family- and community-level regulatory mechanisms that are more prominent in Confucian-influenced East Asian contexts (Liu & Travers, 2018).

Situational Catalysts ('Psychological Push and Environment'). The online environment itself acts as a powerful catalyst. Anonymity and deindividuation reduce feelings of accountability and weaken internal inhibitions, enabling behaviours that individuals might avoid in face-to-face interactions. This is compounded using neutralisation techniques, in which perpetrators employ cognitive rationalisations (e.g., 'it was just a joke,' 'they deserved it') to morally disengage and justify their actions (Sykes & Matza, 1957). The digital context thus provides the 'psychological push and environment' conducive to offending.

Opportunity Structure ('Feasibility Conditions'). For cyberbullying to occur, corresponding opportunities must exist. Routine Activity Theory identifies three necessary elements: a motivated offender, a suitable target, and the absence of a capable guardian (Lian et al., 2023; Ige & Adewale, 2022; Cohen & Felson, 1979). In cyberspace, potential victims who may be vulnerable to online harm are highly accessible ('suitable targets'), and the lack of effective supervision or immediate intervention from parents, platforms, or peers ('absence of guardians') creates the 'feasibility conditions' for harm to unfold. Below is a flowchart illustrating the framework informed by the theories (Figure 1).

Figure 1. Flowchart for the Theories Understanding Cyberbullying Perpetration



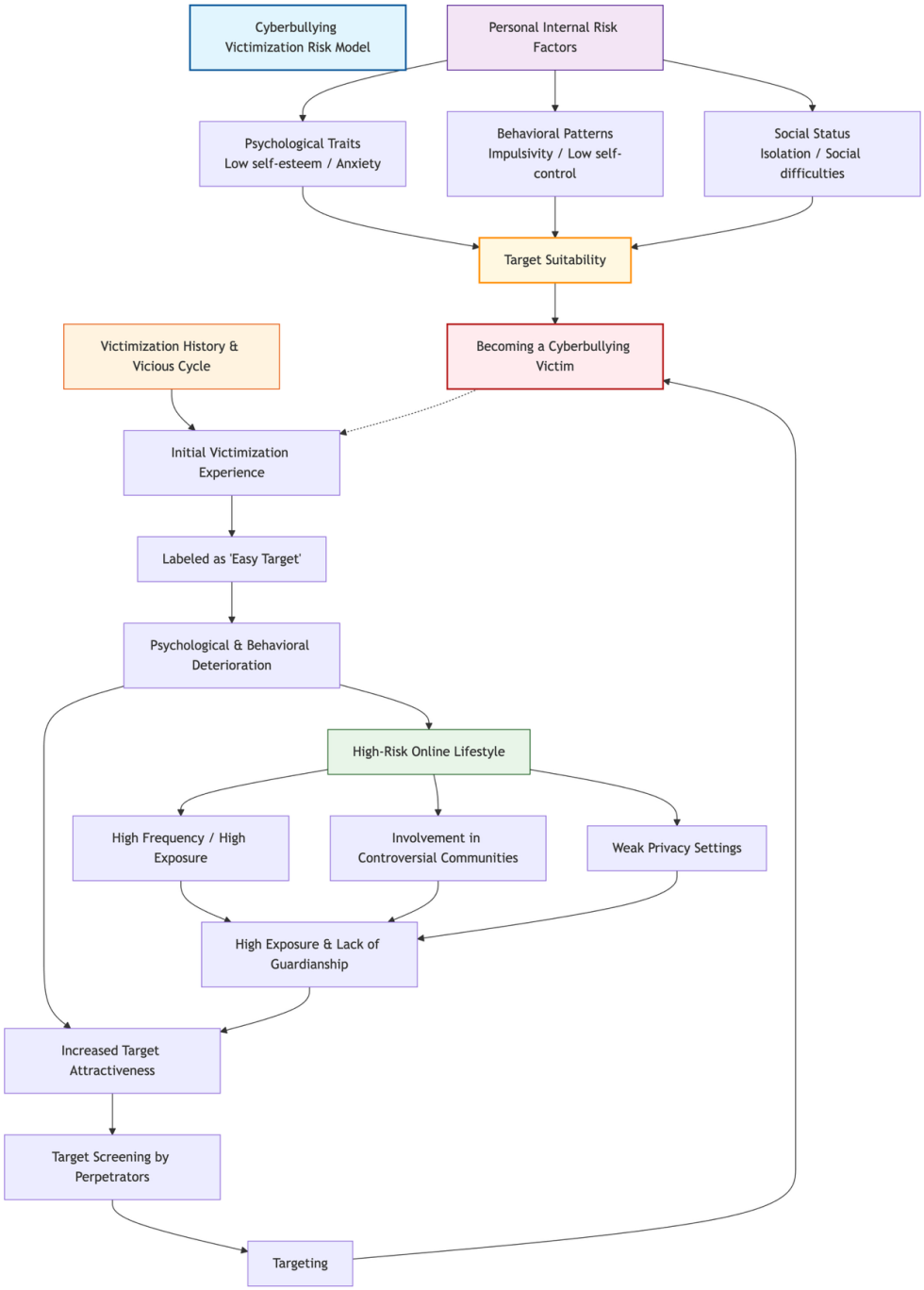
Victim Risk Factors: The 'Target Suitability'

A separate but related line of criminological inquiry examines why certain individuals are at higher risk of cyberbullying victimization. Individual internal risk factors, such as pre-existing psychological traits (e.g., anxiety, low

self-esteem), behavioral patterns (e.g., impulsivity), and social difficulties (e.g., isolation), can increase cyberbullying victimization vulnerability (Kasturiratna et al., 2025; Anichitoae et al., 2025). Furthermore, Henson and Fisher (2011) have proposed cyberlifestyle-routine activity theory (CLRAT) to explain cyber-interpersonal victimization such as cyber harassment and cyber impersonation. According to CLRAT, adolescents' excessive exposure to online risk, proximity to potential online offenders (e.g., engagement in hostile communities), target suitability (linked with the previous 'vulnerability model'), and weak guardianship (e.g., poor privacy management) in cyberspace are common predictors of one or more types of cybervictimization and cyberbullying (Li et al., 2024; Vakhitova et al., 2019).

Crucially, an initial victimization experience can trigger a vicious cycle: the victim may be labeled an 'easy target,' leading to psychological deterioration that further increases their perceived 'target suitability' and exposure, thereby reinforcing future risk. Specifically, when young people experience harassment or aggression in online environments leading to distress, depression, or a reduced sense of well-being, rather than resolving these emotions through adaptive coping mechanisms, some adolescents may turn to the internet as a means of distraction or emotional relief, increasing their risk for further internet-related victimization (Brighi et al., 2019; Chu et al., 2023). Below is the flowchart to summarize the relevant theories regarding cyberbullying victimization (Figure 2).

Figure 2. Flowchart for the Theories Understanding Cyberbullying Victimization



Methodology: A Systematic Review with Thematic Synthesis

This study employed a systematic review methodology following PRISMA 2020 guidelines (Page et al., 2021), designed to comprehensively synthesise the heterogeneous body of literature on AI chatbots for cyberbullying. Given the diversity of relevant studies encompassing technical evaluations, user-centred design research, and theoretical papers, we employed rigorous systematic processes for search, screening, and data extraction, followed by thematic synthesis of findings.

Research Design and Philosophical Approach

The review was framed within a qualitative integrative analysis paradigm, designed to synthesize knowledge across methodological boundaries. This approach is particularly suited for emerging, interdisciplinary fields where evidence is heterogeneous, encompassing both quantitative performance metrics and qualitative insights into user experience and design. The synthesis was informed by a socio-technical criminological lens, recognizing that the effectiveness of AI interventions cannot be assessed through technical capabilities alone but must be understood within the complex social, psychological, and ethical contexts of adolescent online life. Foundational criminological and victimological theories, including the General Theory of Crime, Social Learning Theory, Routine Activity Theory, and the Cyberbullying Victimisation Risk Model, provided an analytical framework for interpreting how chatbot functions interact with the mechanisms of online offending and victimisation.

Search Strategy and Information Sources

A systematic, iterative search strategy was conducted across multiple disciplines, utilising primary academic databases to ensure comprehensive coverage.

Databases Searched:

A comprehensive literature search was conducted spanning from January 2018 to July 2025, with the final search executed on July 15, 2025. To ensure thorough coverage, the search encompassed interdisciplinary databases (Scopus, Web of Science Core Collection) for broad citation tracking, computer science and engineering databases (IEEE Xplore, ACM Digital Library) for technical AI and NLP literature, and psychology and education databases (PubMed, PsycINFO, ERIC) for behavioral interventions and educational applications. This was supplemented by reviewing the first 200 results from Google Scholar and by manual forward and backward citation chaining. The search strategy was constructed by combining keywords and Boolean operators across three conceptual clusters: technology ("AI chatbot" OR "conversational agent" OR "virtual assistant" OR "dialogue system" OR ("artificial intelligence" AND "chat")), the problem domain ("cyberbully*" OR "cyber bully*" OR "online harass*" OR "digital aggression" OR "online victimi*"), and the context or function (detect* OR prevent* OR interven* OR support OR educat* OR "victim*" OR "bystander*").

Example Full Search String (Scopus):

> TITLE-ABS-KEY(("AI chatbot*" OR "conversational agent*" OR "virtual assistant*" OR "dialogue system*" OR ("artificial intelligence" AND "chat*")) AND (cyberbully* OR "cyber bully*" OR "online harass*" OR "digital aggression" OR "online victimi*") AND (detect* OR prevent* OR interven* OR support OR educat* OR "victim*" OR "bystander*")) AND PUBYEAR > 2017 AND PUBYEAR < 2026

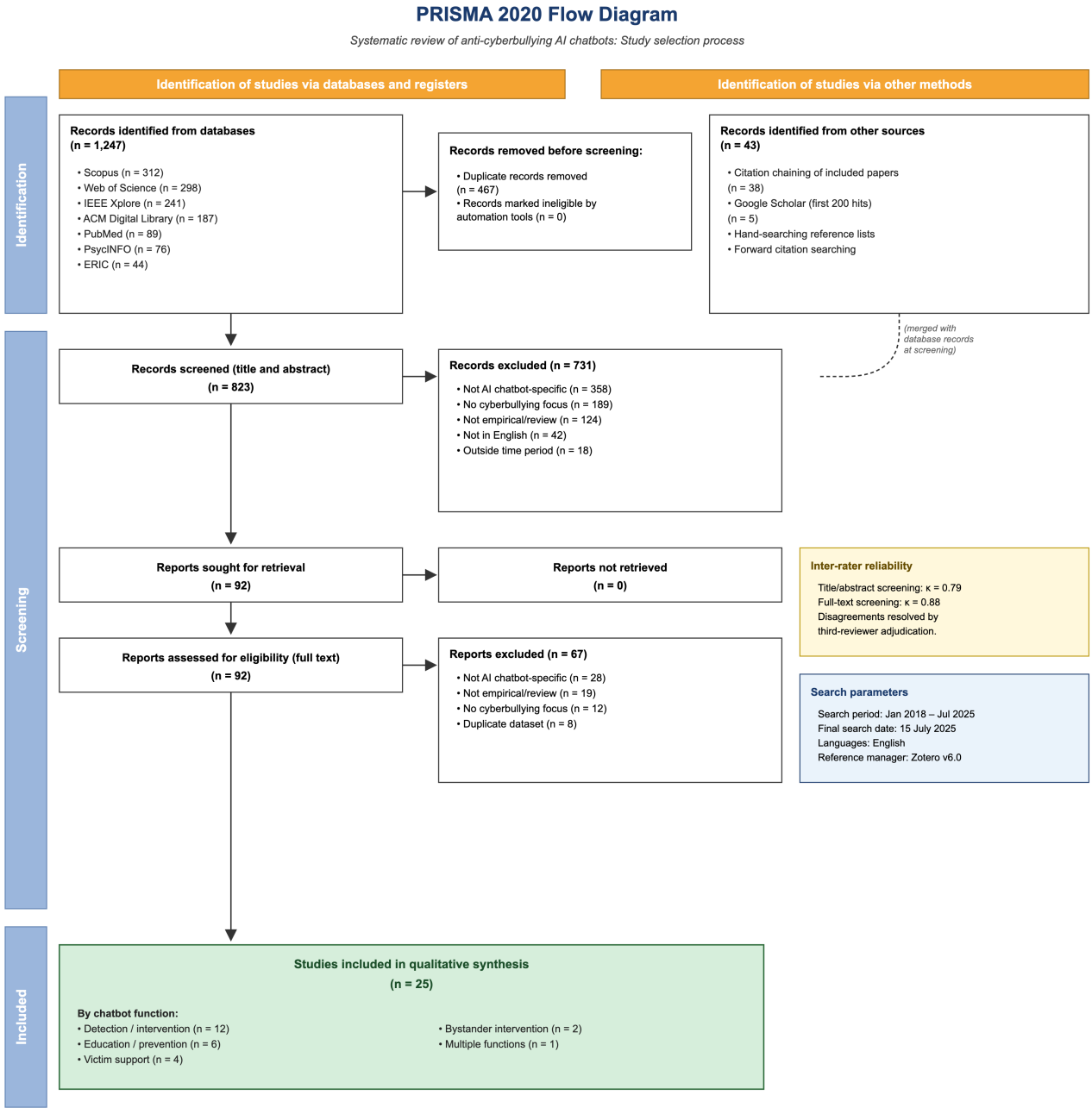
Eligibility Criteria - Inclusion/ Exclusion Criteria:

To be eligible for inclusion in the review, literature had to consist of peer-reviewed journal articles or conference papers presenting empirical studies (quantitative, qualitative, or mixed methods) or systematic reviews. Eligible studies were required to have an explicit focus on the design, implementation, or evaluation of AI-driven chatbot systems targeting cyberbullying or online harassment in the context of prevention, detection, intervention, victim support, or bystander education. Additionally, included works had to be published in English between January 2018 and July 2025, with their full texts accessible. Conversely, articles were excluded if they focused on general AI content moderation without a specific chatbot interface, or if they solely addressed traditional (offline) bullying without a digital component. Further exclusion criteria encompassed non-peer-reviewed opinion pieces or editorials lacking empirical data, duplicate reports of the same dataset, and studies where the full text remained inaccessible despite exhaustive retrieval attempts.

Study Selection and Screening

Following PRISMA guidelines, the study selection process began with deduplication in Zotero using DOI matching, title matching, and a manual review of uncertain cases. The resulting pool of unique records then underwent a two-stage screening process. In the first stage, two researchers independently screened all titles and abstracts against the predefined eligibility criteria. Any disagreements were resolved through discussion, with a third researcher consulted for unresolved cases. In the second stage, the two researchers independently evaluated the full-text articles for final eligibility, resolving any remaining disagreements through consensus. The complete selection process is documented in the PRISMA Flow Diagram (Figure 3).

Figure 3. PRISMA flow diagram



Adapted from: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi:10.1136/bmj.n71. For more information visit: http://www.prisma-statement.org/

Note. Records identified through databases (n = 1,247) and other sources (n = 43) were combined and de-duplicated prior to title/abstract screening (1,290 − 467 duplicates = 823 unique records). The merged-stream approach is reflected by the dashed arrow joining the two identification paths at the screening stage.

Data Extraction

Data extraction was conducted using a standardized form designed to capture a comprehensive range of study details. This included general study characteristics (authors, year, country, publication venue) and specific chatbot features (purpose, role, target users, platform). The form also recorded the theoretical frameworks and criminological theories utilized, alongside technical methods such as AI/ML techniques, NLP approaches, and model architectures. Additionally, dataset characteristics, evaluation approaches, performance metrics (e.g., accuracy, precision, recall, F1), and key findings were documented. To ensure reliability, two researchers independently extracted data from a random 30% of the included studies. A single researcher completed the extractions for the remaining studies, with a second researcher verifying the accuracy of a random 20% sample.

Quality Appraisal

The 25 included studies were published between 2018 and 2025 (median publication year = 2023). Quality appraisal scores derived from criteria adapted from MMAT (Hong et al., 2018) and JBI Critical Appraisal Tools ranged from 6 to 13 out of 14 (mean = 9.7, SD = 1.9), indicating moderate overall methodological quality. Two researchers independently appraised all 25 included studies, achieving high inter-rater reliability, and resolved any minor disagreements through discussion. To make the implications of these scores transparent, Table 1 translates the numeric ratings into four evidentiary weight categories: High, Moderate, Low, and Indicative (proof-of-concept) based on source type, empirical design, sampling rigour, independence, and real-world generalizability.

Data Synthesis

Due to the methodological heterogeneity of the included studies, a narrative synthesis was conducted rather than a statistical meta-analysis (Popay et al., 2006). The synthesis process began with an immersive familiarization of the extracted data, followed by open coding to identify recurring concepts and group them into descriptive themes. These themes were then analysed using criminological frameworks. Finally, a narrative synthesis was developed to compare findings and explore contradictions. To directly address the research questions, the synthesized findings were organized by the chatbot's role-specifically focusing on detection, victim support, education, and bystander intervention and analyzed through the lens of established criminological theories.

Table 1. Formal Quality Appraisal and Evidentiary Weight of Included Studies

Study	Source Type & Venue	Empirical Design	Sample / Validation Rigor	Independence of Evaluation	Real-World Generalizability	Overall Evidentiary Weight
Milosevic et al. (2023)	Peer-reviewed journal article (Social Media + Society, Q1, SSCI)	Qualitative - semi-structured interviews + focus groups	59 adolescents (12-17), Ireland; thematic saturation reported; sample homogeneity acknowledged	Independent academic team; no commercial conflict	Moderate - single national context but methodologically transparent	High
Zou et al. (2024)	Peer-reviewed journal article (Int. J. Child-Computer Interaction, Q1)	Qualitative participatory design (2-stage) + Wizard-of-Oz prototyping	17 participants (10 teens, 7 educators); thematic analysis with open/axial coding	Independent academic team	Moderate - well-grounded design parameters but no field deployment	High
Maenhout et al. (2021)	Peer-reviewed journal article (Frontiers in Public Health, Q1)	Mixed-methods (Person-Based Approach); pilot test with log data + questionnaires + interviews	36 + 6 + 73 + 13 across 3 stages; convenience sampling acknowledged	Independent academic team	Moderate-to-High - actual pilot deployment in real-life setting	High
Kumar et al. (2024)	Peer-reviewed journal article (Electronics, MDPI, Q2)	Quantitative comparative ML benchmarking + synthetic data generation	~48,000 sentences; transparent reporting of validation loss and overfitting flags; AUC not reported	Independent academic team	Low-to-Moderate - English Twitter only; no deployment	Moderate
Muneer & Fati (2020)	Peer-reviewed journal article (Future Internet, MDPI, Q2)	Quantitative comparative ML benchmarking	37,373 tweets; standard 70/30 split; complete metric reporting (acc, prec, rec, F1, time)	Independent academic team	Low-to-Moderate - English benchmark only	Moderate
Lee, Lee & Lee (2022)	Working paper / preprint (SSRN); not peer-reviewed at time of citation	Quantitative experimental survey (moderated mediation)	303 South Korean adults (nonclinical); single 5-min interaction; self-report only; no control group	Independent academic team	Low - short, non-clinical interaction; survey-based	Moderate (caution: pre-print)
Hedderich et al. (2024)	Peer-reviewed conference paper (CHI '24, top-tier HCI venue)	Qualitative design probe study (think-aloud + interviews)	13 middle-school teachers; affinity diagramming	Independent academic team	Low-to-Moderate - design implications only; no student outcomes	Moderate
Piccolo, Troullinou & Alani (2021)	Peer-reviewed conference paper (DIS '21, reputable HCI venue)	Qualitative participatory design (LEGO performance method)	110 UK schoolchildren (11-17); single session	Independent academic team	Low-to-Moderate - UK-specific; no functional prototype tested	Moderate

Study	Source Type & Venue	Empirical Design	Sample / Validation Rigor	Independence of Evaluation	Real-World Generalizability	Overall Evidentiary Weight
Lian, Costilla Reyes & Hu (2023)	Peer-reviewed conference paper (HCII '23, LNAI Springer)	Quantitative ML benchmark + scenario-based functional demo	71,350 messages from public datasets; no user testing; developer-run scenarios only	Independent academic team	Low - proof-of-concept only; authors acknowledge absence of evaluation with target population	Indicative (proof-of-concept)
Sidhu & Sidhu (2025)	Book chapter (Springer edited volume); systematic review	Narrative/conceptual synthesis of 14 reviews	Search transparent (Scopus, PsycInfo, PsycArticles); no primary data	Independent academic team	Low - synthesizes secondary claims; framework not tested	Moderate (as a review)
Zahroh, Kristanto & Dewi (2025)	Peer-reviewed journal article (Jurnal Teknologi Pendidikan, regional Indonesian journal)	Literature review with thematic synthesis	Search strategy described but not fully systematic; no quality appraisal of included studies	Independent academic team	Low - no primary data; framework is conceptual	Low-to-Moderate
Mendoza-Pinto (2025)	Peer-reviewed journal article (CLEI Electronic Journal, regional Latin American journal)	Exploratory-descriptive technical development + simulated testing	6 simulated scenarios; no real users; subjective ratings	Single-author study	Low - controlled simulation only; no field deployment	Indicative (proof-of-concept)
Mendoza-Pinto (2024)	Journal venue not provided in source excerpt - citation completeness flagged	Technical development + functional verification	No formal dataset; no users	Single-author study	Very low - functional verification only	Indicative (proof-of-concept)
Sanu, Mummigatti & Mohana (2023)	Conference proceedings (ICSCNA 2023)	Rule-based system demonstration; no AI/ML	No dataset; demonstration via screenshots only	Independent academic team	Very low - rule-based prototype with no learning capability	Indicative (proof-of-concept)

Results: Thematic Synthesis of the Literature

Following the systematic search and screening process, 25 studies met the inclusion criteria. The heterogeneous nature of the evidence - spanning technical validations, user studies, participatory design research, and pedagogical frameworks precluded statistical meta-analysis. Instead, we conducted qualitative thematic synthesis supported by two analytical instruments: a formal quality appraisal that assigns each source an explicit evidentiary weight (Table 1), and a disaggregated mapping of performance metrics, datasets, and evaluation settings (Table 2). Together, these tools allow readers to trace each substantive claim back to the empirical conditions that underlie it.

Study Characteristics and Quality Assessment

The distribution of evidentiary weight (in Table 1) is itself a substantive finding. Only three studies (Milosevic et al., 2023; Zou et al., 2024; Maenhout et al., 2021) qualified as High evidentiary weight; each was a peer-reviewed Q1 journal article reporting independently conducted, methodologically transparent qualitative or mixed-methods empirical work, and notably each centred user experience rather than algorithmic performance. A larger group of studies (Kumar et al., 2024; Muneer & Fati, 2020; Hedderich et al., 2024; Piccolo, Troullinou & Alani, 2021; Sidhu & Sidhu, 2025) qualified as Moderate evidentiary weight: peer-reviewed work with clear methods but significant limitations in generalizability, dataset representativeness, or absence of deployment evaluation. The Lee, Lee & Lee (SSRN) preprint was retained at moderate weight only with explicit caution flags. Four studies (Lian et al., 2023; Mendoza-Pinto, 2024, 2025; Sanu et al., 2023) were classified as Indicative (proof-of-concept) because they provided technical demonstrations or simulated scenario testing without empirical validation involving target populations. Zahroh et al. (2025) was rated Low-to-Moderate as a regional review without primary data.

This stratification has direct implications for how readers should interpret subsequent claims. Statements grounded in High-weight evidence (e.g., adolescent design preferences for autonomy and agency) can be advanced with reasonable confidence within their cultural context. Statements grounded in Indicative or Low-weight evidence (e.g., chatbot detection performance figures, ChatGPT-based victim support efficacy) must be presented as preliminary signals requiring further validation rather than established findings. The next section presents the performance metrics for all detection and support studies (see Table 2 below).

Table 2. Performance Metrics

Study	Model(s) Reported	Dataset & Size	Language	Evaluation Setting	Accuracy	Precision	Recall	F1	AUC / Other	Validation Method	Real-World Deployment?
Lian et al. (2023)	Stochastic Gradient Descent (SGD) Classifier (selected from 9 candidates)	71,350 annotated messages (31,300 offensive) compiled from 3 public datasets	English	Laboratory benchmark + developer-run scenario tests (no end-user testing)	89.13%	94.46%	84.01%	88.93%	Not reported	Held-out test set; binary classification only	No - authors explicitly note absence of user testing with adolescents

Study	Model(s) Reported	Dataset & Size	Language	Evaluation Setting	Accuracy	Precision	Recall	F1	AUC / Other	Validation Method	Real-World Deployment?
Kumar et al. (2024)	Transformer suite: DeBERTa, BERT, Longformer, HateBERT, DistilBERT, RoBERTa, MobileBERT, XLNet, ELECTRA, BigBird	~48,000 balanced Twitter/X sentences + LLM-generated synthetic data (jailbroken ChatGPT-4, Pi AI, Claude 3, Gemini-1.5)	English	Laboratory benchmark (multilabel: cyberbullying × bias type)	98.46-98.86% (validation accuracy after 1 epoch)	Not separately reported per class	Not separately reported per class	Not separately reported	Validation loss 0.049-0.073	Train/validation split; some models flagged for high overfitting risk (RoBERTa, MobileBERT)	No - prototype apps only ("CyberBulliedBiasedBot"); no field deployment
Muneer & Fati (2020)	7 traditional ML classifiers (LR, LGBM, SGD, RF, AdaBoost, Multinomial NB, SVM); TF-IDF + Word2Vec features	37,373 unique tweets (global dataset)	English	Laboratory benchmark	LR 90.57%; SGD 90.60%; SVM 67.13%	LR 0.9518; SGD 0.9683; NB 0.7952	LR 0.9053; SVM 1.0000; NB 0.9736	LR 0.9280; SGD 0.9270; SVM 0.8033	Not reported	70/30 train-test split	No - comparative analysis only
Mendoza-Pinto (2025)	ChatGPT API integrated into Telegram bot (rule-based + LLM hybrid)	6 simulated test scenarios (no benchmark dataset)	Spanish	Controlled simulation (no real students)	Not formally computed	Not reported	"90% of relevant terms identified" (self-described)	Not reported	Subjective rating: 85% of responses rated empathetic	Author-conducted scenario testing; no statistical validation	No - controlled lab simulation; future deployment proposed
Mendoza-Pinto (2024)	ChatGPT API + Google Apps Script (OttoBot)	No formal dataset; functional verification only	Spanish	Controlled functionality test	Not reported	Not reported	Not reported	Not reported	Not reported	Developer scenario testing	No - proof-of-concept
Sanu et al. (2023)	Rule-based chatbot; regex keyword matching (no ML)	Hard-coded JSON knowledge base (no training data)	English	Demonstration only	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Functional demonstration via screenshots	No - proof-of-concept
Zahroh et al. (2025)	LLMs (Claude 3.0, Mistral, ChatGPT-4) - cited from Cirillo et al. (2025)	Not specified in the review	English (presumed)	Secondary citation in literature review	"High accuracy" (no figure independently verified)	Not reported	Not reported	Not reported	Not reported	Not independently validated	No - literature review
Sidhu & Sidhu (2025)	Multiple ML/NLP models discussed (SVM, RF, BERT, RNN/LSTM, CNN-VGG16, CNN-InceptionV3)	Aggregated across 14 reviewed papers	Multilingual coverage discussed conceptually; primary studies dominantly English	Systematic review (no primary metrics)	Not reported	Not reported	Not reported	Not reported	Not reported	Narrative synthesis of 14 reviews	No - review article

Disaggregated Performance Metrics: Behind the 89-99% Accuracy Range

Table 2 disaggregates the metrics underlying the 89-99% accuracy range cited in some studies, revealing important qualifications that summary reporting would have obscured. First, the upper bound of this range is derived almost exclusively from a single study (Kumar et al., 2024), which reported validation accuracy of 98.46-98.86% after only one training epoch on a balanced dataset comprising approximately 48,000 Twitter/X sentences, supplemented with synthetic data generated by jailbroken commercial LLMs. The authors themselves flagged several models in their suite (RoBERTa, MobileBERT) as exhibiting high overfitting risk, and no per-class precision, recall, or F1 scores were reported. Validation accuracy on a held-out partition of the training distribution is not equivalent to operational performance and cannot support inferences about real-world deployment effectiveness.

Second, the lower bound of the range (89.13%) comes from Lian et al. (2023), whose SGD classifier achieved 94.46% precision but only 84.01% recall. This asymmetry is criminologically significant: the system would correctly flag most posts it identifies as cyberbullying, but would miss approximately one in six instances of actual harm. In victimological terms, false negatives represent missed protective opportunities - victims whose harm is not detected and who therefore receive no intervention - and the social cost of these errors differs categorically from false positives. The summary metric of "accuracy" masks this asymmetry entirely.

Third, Muneer and Fati (2020) provide the most complete metric reporting in the corpus, with logistic regression achieving 90.57% accuracy, 0.9518 precision, 0.9053 recall, and 0.9280 F1 on a 70/30 split of 37,373 tweets. Yet even this comparatively rigorous study reported no AUC, no class-imbalance-adjusted metrics, and no temporal validation. Their SVM classifier, while achieving perfect recall (1.000), did so at the cost of 67.13% accuracy, illustrating that single-metric reporting can be highly misleading.

Fourth, and most consequentially, none of the proof-of-concept studies (Lian et al., 2023; Mendoza-Pinto, 2024, 2025; Sanu et al., 2023) reported empirical performance measures involving the target population. Mendoza-Pinto (2025), for instance, evaluated a ChatGPT-Telegram bot prototype using six developer-conducted simulated scenarios with self-rated empathy scores, while Sanu et al. (2023) provided only functional screenshots of a rule-based system with no learning capability. Treating these demonstrations as evidence of "chatbot effectiveness" conflates technical feasibility with prevention outcomes.

Across all detection studies, three further structural absences emerge from Table 1. No study reported AUC or any class-imbalance-adjusted metric, despite cyberbullying constituting less than 1% of content on real platforms. No study employed temporal validation (training on earlier data and testing on later data) to assess robustness against the rapid evolution of online slang, memes, and adversarial evasion strategies. No study evaluated systems in real-world deployment with adolescents in naturalistic contexts, nor did it measure downstream criminological outcomes such as reduced victimisation incidence, deterred perpetration, or improved victim well-being. The 89-99% range, therefore,

represents an upper-bound estimate of laboratory classification performance under controlled conditions, not a defensible projection of operational effectiveness.

Geographic and Linguistic Distribution of Evidence

Table 2 shows that all detection benchmarks were conducted on English-language corpora, except for two Spanish-language proof-of-concept demonstrations by Mendoza-Pinto (2024, 2025). No included study evaluated detection performance on Bahasa Indonesia, Bahasa Malaysia, Tagalog, Thai, Vietnamese, Korean, Japanese, or any of the major Chinese variants spoken in the Asia Pacific. Code-switched and multilingual corpora, the linguistic reality of much adolescent communication in the region, were entirely absent from the validated evidence base. Table 1 reinforces this concentration on the methodological side. All three High-weight participatory and mixed-methods studies (Milosevic et al., 2023 - Ireland; Zou et al., 2024 - USA; Maenhout et al., 2021 - Belgium) were conducted in Western individualist contexts. No High-weight study examined design preferences, trust formation, or help-seeking behaviour with Asian youth in their own cultural context. The strongest empirical claims about user-centred design, therefore, derive from cultural contexts characterised by individualist help-seeking norms, low-power-distance authority relationships, and Western privacy paradigms, none of which can be assumed to transfer to collectivist, high-power-distance, or face-saving cultural settings prevalent across much of the Asia Pacific.

Thematic Synthesis Overview

Four overarching themes emerged from the synthesis, each interpreted through the dual lenses of evidentiary weight (Table 1) and disaggregated performance evidence (Table 2).

The Evolving Role of AI - From Reactive Tool to Proactive Crime Prevention Partner:

The literature reveals progressive applications spanning detection, victim support, and proactive education. However, evidentiary weight varies sharply across these domains. Detection claims derive predominantly from Indicative or Moderate-weight benchmarking studies on Western English data (Lian et al., 2023; Kumar et al., 2024; Muneer & Fati, 2020); victim support claims rest on Indicative-weight proof-of-concept demonstrations without longitudinal outcome measurement (Mendoza-Pinto, 2024, 2025; Sanu et al., 2023); proactive education claims draw on a mix of Moderate and High-weight participatory work, but exclusively in Western settings (Hedderich et al., 2024; Zou et al., 2024). The trajectory toward proactive crime prevention is therefore conceptually promising but empirically uneven.

The Centrality of User-Centered and Participatory Design: This theme rests on the strongest evidentiary foundation in the corpus. Three High-weight studies (Milosevic et al., 2023; Zou et al., 2024; Maenhout et al., 2021) consistently identify adolescent demand for autonomy, opt-in/opt-out control, empathetic conversational tone, and self-reliance over peer or authority involvement. However, the geographic concentration in Western contexts means these design principles cannot be straightforwardly generalised; they should be treated as well-evidenced for Western individualist settings and as testable hypotheses for collectivist Asian contexts.

Navigating the Socio-Ethical Landscape: Privacy-safety tensions, trust formation, and chilling effects from false positives are recurrently identified across studies of varying evidentiary weight. The disaggregated false-positive/false-negative analysis (Table 1) provides quantitative grounding for the chilling-effects argument: detection systems with substantially asymmetric error profiles will produce systematically different governance consequences depending on which error type dominates.

Implementation Challenges and Future Directions: Technical limitations in handling slang, sarcasm, code-switching, and culturally specific aggression are noted across the literature, but no study in our corpus empirically evaluated cross-linguistic transfer. The collaboration imperative that AI chatbots succeed only as complementary tools within human-led ecosystems emerges consistently across High-, Moderate-, and Indicative-weight studies, lending it cross-evidentiary robustness.

Discussion: Interpreting the Synthesized Evidence Through a Criminological Lens

Tables 1 and 2 together establish the interpretive frame for this Discussion. The appraisal exercise reveals a sharply stratified evidence base: three High-weight studies grounded in user-centred and participatory design within Western contexts, a larger cluster of Moderate-weight benchmarking studies trained on English-language Western datasets, and a non-trivial group of Indicative proof-of-concept demonstrations whose contribution is technical feasibility rather than demonstrated efficacy. This stratification is not a methodological footnote but the analytical foundation on which every subsequent inference rests. Accordingly, the sections that follow calibrate each interpretive claim to the evidentiary weight of its underlying source, treating findings as inferences about the cultural context in which the evidence was generated rather than as universal propositions. Four threads organize the analysis: the laboratory-to-practice gap in detection performance, the evidence for victim support, the geographic and linguistic concentration of the evidence base, and the stratification of practice implications by evidentiary strength.

The Laboratory-to-Practice Gap, Quantified

The disaggregated metrics in Table 2 give empirical specificity to what is often asserted in general terms as a "laboratory-to-practice gap" in algorithmic cyberbullying detection. Five concrete gaps are visible in the appraised evidence. First, distribution shift: no included study evaluated temporal robustness, yet online language evolves rapidly, and Kumar et al.'s (2024) one-epoch validation provides no defence against this drift. Second, adversarial behaviour: benchmark datasets contain no deliberate evasion attempts, yet real deployment provokes immediate adversarial adaptation through misspellings, character substitutions, and emoji-coded communication. Third, context collapse: the same utterance may be playful banter or aggression depending on relational history, but the studies in Table 1 evaluated isolated text classification without contextual metadata. Fourth, multilingual and code-switched communication: the region's linguistic reality is structurally absent from the validated evidence. Fifth, class imbalance: every reviewed benchmark used balanced or semi-balanced data, whereas real platforms exhibit cyberbullying base rates below 1%. Models that achieve 90% accuracy on balanced sets routinely produce false-positive rates that overwhelm review capacity at realistic base rates.

The criminological implication is that the algorithmic guardianship envisioned under Routine Activity Theory has been demonstrated only in highly controlled, decontextualised conditions. Whether such guardianship operates in field deployment and whether it deters motivated offenders, protects suitable targets, or instead produces over-flagging that erodes legitimacy remains empirically open across all reviewed evidence.

Re-evaluating the Evidence for Victim Support

Claims about chatbot-mediated victim support in this literature rest on a narrower evidentiary base than citation frequency suggests. The Mendoza-Pinto (2024, 2025) studies are appraised as Indicative (proof-of-concept) given their reliance on simulated scenarios, single-author conduct, and absence of empirical user testing. The Lee, Lee and Lee SSRN preprint, although providing experimental data, is assessed as Moderate weight with caution flags due to its preprint status, single-session design, exclusive reliance on self-report, and absence of a control group. No High-weight evidence in the corpus supports the claim that AI chatbots reduce victimisation-related distress, depression, or suicidal ideation over time. The literature on victim support must therefore be read carefully: it documents engagement and acceptability under controlled conditions, not therapeutic effectiveness or longitudinal outcomes consistent with victimological standards. Reading these studies as evidence of efficacy rather than feasibility risks overstating the readiness of chatbot interventions for vulnerable populations whose distress, by definition, exceeds what a single controlled session can capture.

Implications for the Western-Dataset Bias Argument

The geographic and linguistic concentration documented in Table 2 (zero Southeast Asian language studies) and the cultural concentration documented in Table 1 (all High-weight participatory work conducted in Western contexts) jointly substantiate what would otherwise be an inferential claim about applicability gaps. The argument is not that Asia Pacific contexts might differ from Western contexts in ways the evidence has not yet addressed; it is that the evidence base has not empirically addressed Asia Pacific's defining linguistic and cultural features at all. This is a stronger and more defensible position, and it directly grounds the central caution against uncritical technology transfer.

The inference for criminal justice professionals is correspondingly more pointed. Detection systems whose validated performance is confined to English Twitter cannot be assumed to perform comparably on WeChat, LINE, or KakaoTalk; participatory design principles validated with British, Belgian, and American adolescents cannot be assumed to transfer to youth in Confucian-influenced or Islamic-majority Asian contexts. These are not speculative concerns but evidentiary facts established by the appraisal.

Stratifying the Practice Implications by Evidentiary Strength

The five criminological practice implications advanced in this review can now be qualified according to the strength of their underlying evidence. Recommendations for user-agency-centred design, opt-in/opt-out functionality, and empathetic conversational tone draw on High-weight participatory evidence and can be advanced with reasonable confidence within Western contexts and as testable hypotheses elsewhere. Recommendations involving structured referral pathways, detection-based triage, and AI-mediated victim support draw predominantly on Indicative or Moderate-weight evidence and must be presented as starting hypotheses rather than validated practices. Recommendations for capacity-building among criminal justice professionals are conceptual and cross-cutting, defensible on procedural justice grounds independent of any specific empirical study. Recommendations for culturally adapted protocols and co-design with Asian youth follow from the absence of relevant evidence rather than from positive findings; they identify research priorities as much as practice directions.

Synthesising the Position

Read together, Tables 1 and 2 reframe the discussion from a balanced summary of what AI chatbots can do into a more disciplined account of what the evidence currently licenses us to conclude. The headline 89-99% accuracy figure, while not incorrect as a literal description of laboratory performance, conveys an unwarranted impression of operational readiness once disaggregated. The user-centred design literature is more empirically robust than aggregated summaries suggest, but it is also more geographically narrow. The victim support literature is weaker than its citation frequency suggests. The proactive education literature is conceptually promising but resource-bounded. Across all four themes, the absence of validated evidence from Asia Pacific languages, platforms, and cultural contexts is not a peripheral limitation but a defining characteristic of the field.

For criminological practice in the Asia Pacific, the central conclusion is therefore neither optimistic nor dismissive but disciplined. AI chatbots warrant inclusion in the toolkit of crime prevention and victim support, but only as one element among several: supported by local data, validated through deployment research, governed by ethical safeguards developed with affected communities, and integrated with the human relationships and cultural expertise that the algorithmic literature cannot replace. The path forward is not faster algorithmic deployment but more rigorous local research, more transparent metric reporting, and more sustained interdisciplinary collaboration between criminologists, technologists, and the youth populations whose lives these systems will shape.

Conclusion: Towards a Disciplined, Locally Grounded Future

This systematic review demonstrates that AI chatbots are neither a single solution nor a transformative technology, but a stratified set of tools whose claimed efficacy depends heavily on the evidentiary conditions under which they have been demonstrated. The disaggregated performance evidence (Table 2) and formal quality appraisal (Table 1) jointly reframe the field: a small number of High-weight studies establish reasonably confident findings about user-centred design preferences within Western individualist contexts; a larger group of Moderate-weight benchmarking studies establish laboratory classification performance on English-language Western datasets; and a non-trivial proportion of the literature consists of Indicative proof-of-concept demonstrations that document technical feasibility rather than prevention outcomes.

The criminological frameworks invoked in this review remain analytically useful but must be applied with the corresponding discipline. Detection systems operationalize Routine Activity Theory through algorithmic guardianship, but only under controlled conditions that have not been replicated in the multilingual, code-switched, platform-diverse environments where Asia Pacific youth actually communicate. Role-play tools engage Social Learning Theory, but the supporting evidence derives from high-resource Western educational settings whose pedagogical traditions, authority structures, and digital infrastructure differ markedly from those in much of the region. Resilience-building interventions address the Victimisation Risk Model factors, but no longitudinal study in our corpus has measured whether such interventions reduce victimisation incidence, distress, or revictimisation over time.

The evidence base has six structural limitations that constrain the inferences this review can support. First, all detection benchmarks were conducted on English-language Western datasets; the only non-English exceptions were Spanish-language proof-of-concept demonstrations. Second, no included study evaluated downstream criminological outcomes such as reduced cyberbullying incidence, deterred perpetration, or improved victim well-being. Third, no study reported AUC or class-imbalance-adjusted metrics, and only a minority disaggregated precision, recall, and F1 alongside accuracy. Fourth, no study employed temporal validation or real-world deployment evaluation with adolescents in naturalistic settings. Fifth, all High-weight participatory design research was conducted in Western individualist contexts. Sixth, no study examined Asia Pacific platform ecologies (WeChat, LINE, KakaoTalk) or Southeast Asian languages. The 89-99% accuracy range, therefore, represents an upper-bound estimate of laboratory performance under controlled conditions, not an empirical projection of operational effectiveness in Asia Pacific deployment contexts.

Drawing on Asian criminology (Liu & Travers, 2018, 2019) and Southern criminology (Carrington et al., 2016, 2019), the central conclusion is that the field stands at a decision point between two trajectories. The first is uncritical technology transfer, in which Western-validated systems are deployed in Asia Pacific contexts on the strength of laboratory benchmarks and Western user studies, with predictable risks: discriminatory false positives against minority linguistic groups, failure to detect culture-specific forms of harm such as honour-based shaming or face-loss aggression, erosion of trust in criminal justice institutions, and exacerbation of existing inequalities between privileged urban youth and rural or economically disadvantaged populations. The second is a disciplined, locally grounded trajectory in which the region's linguistic diversity, technological dynamism, and rich criminological traditions become the foundation for region-specific detection models, participatory design with Asian youth, deployment validation in local settings, and outcome measurement consistent with criminological standards.

For Asia Pacific criminal justice professionals, three priorities follow from this disciplined position. First, build robust safeguards: ethical guardrails, transparent data governance, accountability measures, and ongoing bias auditing developed through consultation with affected communities rather than imposed by external technical standards. Second, invest in local research and development: region-specific detection models trained on local-language and code-switched datasets; participatory design with Asian youth; validation in Asia Pacific platform ecologies; and outcome measurement linked to victimological and criminological indicators rather than classification metrics alone. Third, centre youth as co-creators, consistent with children's rights frameworks and procedural justice principles.

AI chatbots warrant inclusion in the toolkit of crime prevention and victim support in the Asia Pacific, but only as one element among several, supported by local evidence, governed by culturally appropriate ethics, and integrated with the human relationships and cultural expertise that algorithmic systems cannot replace. The path forward is not faster deployment of Western technologies but more rigorous local research, more transparent reporting, and more sustained interdisciplinary collaboration. Without such grounding, AI chatbots risk becoming another instance of inappropriate technology transfer that fails to serve the populations they purport to protect, or worse, harms them. With it, they may become genuinely useful components of locally grounded, ethically defensible, victim-centred criminological practice.

References

[1] Agnew, R. (1992). Foundation for a general strain theory of crime and delinquency. Criminology, 30(1), 47-87.

[2] Anichitoae, F. M., Dobrean, A., Georgescu, R. D., & Roman, G. D. (2025). Association between self-related cognitions and cyberbullying victimization in children and adolescents. Aggression and Violent Behavior, 80, Article 102021. https://doi.org/10.1016/j.avb.2024.102021

[3] Bandura, A. (1978). Social learning theory of aggression. Journal of Communication, 28(3), 12-29.

[4] Barlett, C. P., & Gentile, D. A. (2022). Attacking others online: The formation of cyberbullying in late adolescence. Psychology of Popular Media, 1(2), 123-135. https://doi.org/10.1037/a0028113

[5] Bauer, R. (2025, October 28). The future of youth mental health in the age of AI: Insights from JED's 2025 policy summit. The Jed Foundation. https://jedfoundation.org/the-future-of-youth-mental-health-in-the-age-of-ai-insights-from-jeds-2025-policy-summit/

[6] Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczynski, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. Aggressive Behavior, 47(4), 385-395. https://doi.org/10.1002/ab.21948

[7] Brighi, A., Menin, D., Skrzypiec, G., & Guarini, A. (2019). Young, bullying, and connected: Common pathways to cyberbullying and problematic Internet use in adolescence. Frontiers in Psychology, 10, Article 1467. https://doi.org/10.3389/fpsyg.2019.01467

[8] Carrington, K., Hogg, R., & Sozzo, M. (2016). Southern Criminology. British Journal of Criminology, 56(1), 1-20.

[9] Carrington, K., Hogg, R., Scott, J., & Sozzo, M. (2019). The Palgrave Handbook of Criminology and the Global South. Palgrave Macmillan.

[10] Chan, H. C. (O.), & Wong, D. S. W. (2020). The overlap between cyberbullying perpetration and victimisation: Exploring the psychosocial characteristics of Hong Kong adolescents. Asia Pacific Journal of Social Work and Development, 30(3), 164-180. https://doi.org/10.1080/02185385.2020.1761436

[11] Chaudhary, P. K., Yalamati, S., Palakurti, N. R., Alam, N., Kolasani, S., & Whig, P. (2024). Detecting and preventing child cyberbullying using generative artificial intelligence. In 2024 Asia Pacific Conference on Innovation in Technology (APCIT). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/APCIT62007.2024.10673710

[12] Chen, J.-K., & Chen, L.-M. (2020). Cyberbullying among adolescents in Taiwan, Hong Kong, and mainland China: A cross-national study in Chinese societies. Asia Pacific Journal of Social Work and Development, 30(3), 227-241. https://doi.org/10.1080/02185385.2020.1788978

[13] Chu, X., Li, Q., Fan, C., & Jia, Y. (2023). Life stress and cyberbullying: Examining the mediating roles of expressive suppression and online disinhibition. Journal of Youth and Adolescence, 52(8), 1647-1661. https://doi.org/10.1007/s10964-023-01791-w

[14] Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. American Sociological Review, 44(4), 588-608.

[15] David, O. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with artificial intelligence: Current trends and future prospects. Journal of Medicine, Surgery, and Public Health, 3, Article 100099. https://doi.org/10.1016/j.glmedi.2024.100099

[16] Faraz, A., Ahsan, F., Mounsef, J., Karamitsos, I., & Kanavos, A. (2024). Enhancing child safety in online gaming: The development and application of Protectbot, an AI-powered chatbot framework. Information, 15(4), Article 233. https://doi.org/10.3390/info15040233

[17] Ferrer, R., Ali, K., & Hughes, C. (2024). Using AI-based virtual companions to assist adolescents with autism in recognizing and addressing cyberbullying. Sensors, 24(12), Article 3875. https://doi.org/10.3390/s24123875

[18] Gottfredson, M. R., & Hirschi, T. (1990). A general theory of crime. Stanford University Press.

[19] Guo, Z., Wang, P., Huang, L., & Cho, J.-H. (2023). Authentic dialogue generation to improve youth's awareness of cybergrooming for online safety. In Proceedings of the 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI). Institute of Electrical and Electronics Engineers.

[20] Hedderich, M. A., Bazarova, N. N., Zou, W., Shim, R., Ma, X., & Yang, Q. (2024). A piece of theatre: Investigating how teachers design LLM chatbots to assist adolescent cyberbullying education. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Article 552, pp. 1-17). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642379

[21] Henry, N., Witt, A., & Vasil, S. (2025). A "design justice" approach to developing digital tools for addressing gender-based violence: Exploring the possibilities and limits of feminist chatbots. Information, Communication & Society, 28(11), 1884-1907. https://doi.org/10.1080/1369118X.2024.2363900

[22] Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M. P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M. C., & Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada

[23] Ige, T., & Adewale, S. (2022). AI-powered anti-cyber bullying system using machine learning algorithm of multinomial naive Bayes and optimized linear support vector machine. International Journal of Advanced Computer Science and Applications, 13(5). https://doi.org/10.14569/IJACSA.2022.0130502

[24] Islam, M. S., Sutton, S., & Rafiq, R. (2024). A generative AI-powered approach to cyberbullying detection. In Proceedings of the 2024 8th International Conference on Information System and Data Mining (pp. 80-87). Association for Computing Machinery.

[25] Kasturiratna, K. T. A. S., Hartanto, A., Chen, C. H. Y., Ong, W. H., & Tong, E. M. W. (2025). Umbrella review of meta-analyses on the risk factors, protective factors, consequences and interventions of cyberbullying victimization. Nature Human Behaviour, 9, 101-132. https://doi.org/10.1038/s41562-024-02011-6

[26] Kolomiets, A., Kolomiets, D. L., Kushnir, O. Y., & Tkachyshyn, A. (2025). Possibilities of using artificial intelligence by adolescents in countering cyberbullying. Scientific Issues of Vinnytsia Mykhailo Kotsiubynskyi State Pedagogical University. Section: Pedagogics and Psychology, (81), 37-44. https://doi.org/10.31652/2415-7872-2025-81-37-44

[27] Kumar, P. (2024). Large language models (LLMs): Survey, technical frameworks, and future challenges. Artificial Intelligence Review, 57, Article 260. https://doi.org/10.1007/s10462-024-10888-y

[28] Kumar, Y., Huang, K., Perez, A., Yang, G., Li, J. J., Morreale, P., Kruger, D., & Jiang, R. (2024). Bias and Cyberbullying Detection and Data Generation Using Transformer Artificial Intelligence Models and Top Large Language Models. Electronics, 13(17), 3431. https://doi.org/10.3390/electronics13173431

[29] Li, J. C. M., Jia, C. X., & Mlyakado, B. P. (2024). Assessing online sexual exploitation among secondary school students in Tanzania from a routine activity theory perspective. Child Abuse & Neglect, 147, Article 106597. https://doi.org/10.1016/j.chiabu.2023.106597

[30] Lian, A. T., Costilla Reyes, A., & Hu, X. (2023). CAPTAIN: An AI-based chatbot for cyberbullying prevention and intervention. In Interacción 2023: Proceedings of the XXIII International Conference on Human Computer Interaction. Association for Computing Machinery. https://doi.org/10.1007/978-3-031-35894-4_7

[31] Liu, J., & Travers, M. (2019). Theories of Asian crime and justice. Asian Journal of Criminology, 14(4), 251-256.

[32] Liu, J., & Travers, M. (Eds.). (2018). Comparative Criminology in Asia. Springer.

[33] Marshall, N. J., Loades, M. E., Jacobs, C., West, R. M., & Gamble, L. (2025). Integrating artificial intelligence in youth mental health care: Advances, challenges, and future directions. Current Treatment Options in Psychiatry, 12, Article 11. https://doi.org/10.1007/s40501-025-00348-x

[34] Mathew, A., Sivdutt, S., & B. G. (2025). Prevention of cyber bullying in social media using generative AI. In 2025 6th International Conference on Computing and Data Science (CDS). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICCDS64403.2025.11209751

[35] Mendoza Pinto, R. (2023). Artificial intelligence in the fight against bullying: Integration of ChatGPT in an emotional support chatbot. In CISETC 2023: Congreso Internacional de Sistemas y Tecnologías de la Computación.

[36] Mendoza Pinto, R. (2025). Exploring the potential of ChatGPT in developing a virtual assistant for prevention and support in cases of school bullying. CLEI Electronic Journal, 28(4). https://doi.org/10.19153/cleiej.28.4.6

[37] Milosevic, T., Verma, K., Vigil, S., Carter, M., Staksrud, E., Davis, B., & O'Higgins Norman, J. (2023). Leveraging AI-based interventions to address cyberbullying among children: A rights-based perspective. AoIR Selected Papers of Internet Research. https://doi.org/10.5210/spir.v2022i0.13054

[38] Muneer, A., & Fati, S. M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. Future Internet, 12(11), 187. https://doi.org/10.3390/fi12110187

[39] Naga, Y., Anuradha, H. G., Lalitha, V., & Pravardhitha, N. M. S. (2023). An AI-based mental health support chatbot for cyber bullied victims. In 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). Institute of Electrical and Electronics Engineers.

[40] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71

[41] Piccolo, L. S. G., Troullinou, P., & Alani, H. (2021). Chatbots to support children in coping with online threats: Socio-technical requirements. In DIS '21: Proceedings of the 2021 ACM Designing Interactive Systems Conference (pp. 1504-1517). Association for Computing Machinery. https://doi.org/10.1145/3461778.3462114

[42] Pimpista, D. M., Antunes, A. M., de Almeida, S., Paiva, P. A., da Costa, N., & Ferreira. (2020). Com@Viver: Using affective AI agents to encourage prosocial activity. In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20) (pp. 2112-2114). International Foundation for Autonomous Agents and Multiagent Systems.

[43] Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. ESRC Methods Programme.

[44] Sidhu, M. S., & Sidhu, K. K. (2025). AI strategies for handling disciplinary and cyber bullying in schools. In M. A. Al Sharafi, M. Al Emran, & K. Shaalan (Eds.), Current and future trends on AI applications (Vol. 1178, pp. 207-229). Springer. https://doi.org/10.1007/978-3-031-75091-5_12

[45] Sivadarshini, N., Manickam, V. R., Varshini, R. S., & Baskaran, S. (2025). Cyberbullying detection in heterogeneous data streams using Pytesseract OCR and HateBERT: A cross-modal approach for text, image, and emoji interpretation. International Journal for Research in Applied Science and Engineering Technology, 13(6), 1345-1353. https://doi.org/10.22214/ijraset.2025.69254

[46] Spytska, L. (2025). The use of artificial intelligence in psychotherapy: Development of intelligent therapeutic systems. BMC Psychology, 13(1), Article 175. https://doi.org/10.1186/s40359-024-02294-y

[47] St Martin, L. I. L., & Villeneuve, S. (2024). The uses of chatbots in the context of children and teenagers' bullying: A systematic literature review. Cogent Education, 11(1), Article 2312032. https://doi.org/10.1080/2331186X.2024.2312032

[48] Sulikeri, N. B., Hegde, G. V., Krishna, K. V., Reddy, D. K., Manishimha, G., & Singh, A. (2025). Billy Buddy against cyberbullying. International Journal for Research in Applied Science and Engineering Technology, 13(6), 1685-1692. https://doi.org/10.22214/ijraset.2025.66441

[49] Sykes, G. M., & Matza, D. (1957). Techniques of neutralization: A theory of delinquency. American Sociological Review, 22(6), 664-670.

[50] Ueda, T., Nakanishi, J., Kuramoto, I., Baba, J., Yoshikawa, Y., & Ishiguro, H. (2021). Cyberbullying mitigation by a proxy persuasion of a chat member hijacked by a chatbot. In HAI '21: Proceedings of the 9th International Conference on Human-Agent Interaction (pp. 301-305). Association for Computing Machinery. https://doi.org/10.1145/3472307.3484177

[51] Vakhitova, Z. I., Alston-Knox, C. L., Reynald, D. M., Townsley, M. K., & Webster, J. L. (2019). Lifestyles and routine activities: Do they enable different types of cyber abuse? Computers in Human Behavior, 101, 225-237. https://doi.org/10.1016/j.chb.2019.07.012

[52] Wang, P., Guo, Z., Huang, L., & Cho, J.-H. (2021). SERI: Generative chatbot framework for cybergrooming prevention. In Proceedings of the 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 1153-1158). Institute of Electrical and Electronics Engineers.

[53] Younes, S., Ramzi, D., Osman, S., & Ahmed, M. (2023). Automated detection and response to cyberbullying using machine learning. In 2023 International Conference on Digital Applications, Transformation & Economy (ICDATE). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICDATE58146.2023.10248567

[54] Zahroh, M., Kristanto, A., & Dewi, U. (2025). How can AI-enhanced case-based learning improve problem solving in cyberbullying education?: A literature review. Jurnal Teknologi Pendidikan: Jurnal Penelitian dan Pengembangan Pembelajaran, 10(2), 200-213. https://doi.org/10.33394/jtp.v10i2.14704

[55] Zou, W., Yang, Q., DiFranzo, D., Chen, M., Hui, W., & Bazarova, N. N. (2024). Social media co-pilot: Designing a chatbot with teens and educators to combat cyberbullying. International Journal of Child-Computer Interaction, 41, Article 100680. https://doi.org/10.1016/j.ijcci.2024.100680