applied sciences

MDPI

Review

A Survey of Generative AI for Detecting Pedophilia Crimes

Filipe Silva [1], Rodrigo Rocha Silva [2,3] and Jorge Bernardino [1,2,*]

[1] Coimbra Institute of Engineering (ISEC), Polytechnic University of Coimbra, Rua Pedro Nunes—Quinta da Nora, 3030-199 Coimbra, Portugal; a2022113664@isec.pt
[2] Centre for Informatics and Systems of the University of Coimbra (CISUC), Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal; rrochas@dei.uc.pt
[3] FATEC—Faculdade de Tecnologia de Mogi das Cruzes, São Paulo Technological College, Mogi das Cruzes 08773-600, Brazil
* Correspondence: jorge@isec.pt

Abstract

The complexity for law enforcement and child protection agencies has been exacerbated by the proliferation of child sexual exploitation channels, facilitated by digital platforms and social media. Generative AI's ability to analyze large datasets, recognize patterns, and generate new content makes it one of the potential solutions for detecting suspicious behavior and indicators of child sexual exploitation. This paper discusses the potential of generative AI to aid in the fight against pedophilic crimes by reviewing current research, methodologies, and challenges, as well as future directions and ethical concerns. Although the potential benefits are significant, applying AI to such a sensitive area presents numerous challenges, including privacy concerns, algorithmic bias, and potential misuse, which must be addressed carefully.

Keywords: generative AI; pedophilia crimes; online sexual exploitation; child sexual abuse; victim identification

Academic Editor: Andrea Prati

Received: 7 May 2025
Revised: 21 June 2025
Accepted: 23 June 2025
Published: 24 June 2025

Citation: Silva, F.; Silva, R.R.; Bernardino, J. A Survey of Generative AI for Detecting Pedophilia Crimes. *Appl. Sci.* **2025**, *15*, 7105. https://doi.org/10.3390/app15137105

1. Introduction

Pedophilia is a severe psychosexual disorder characterized by a persistent and often exclusive sexual attraction to children who have not yet reached puberty. Pedophilia causes long-lasting trauma, leading to conditions such as anxiety, depression, PTSD, and difficulty in forming stable relationships [1]. Addressing this problem requires advanced tools, such as AI, to detect and prevent abusive behavior early on.

Online grooming is defined as the process by which an adult manipulates a child through the internet, often with malicious intent, particularly for sexual exploitation. Groomers deploy various tactics such as offering attention, affection, kindness, and gifts to lure and seduce their victims. Core grooming components—including psychological assessment, enticement, cyber exploitation, control, and self-preservation—are critical for identifying grooming behavior. These stages have been extensively studied by Cook et al. (2022) [2], who demonstrated that grooming behavior can be detected through automated systems. However, the nuanced nature of human interactions continues to pose significant challenges.

Unfortunately, the rapid expansion of digital platforms and social media has created new avenues for perpetrators to target and exploit minors. These environments, which provide anonymity and uninterrupted communication, have become breeding grounds for grooming, the dissemination of CSAM, and the orchestration of sexual offenses against

children. The multifaceted role of artificial intelligence in combating child sexual abuse, from identifying CSAM to aiding law enforcement investigations, has been systematically reviewed in recent evidence assessments, underscoring both its potential and the challenges involved in this complex domain [3]. This evolving digital landscape presents formidable challenges to both law enforcement and child protection organizations, who strive to detect and prevent these heinous crimes. As discussed by Levy and Robinson (2022), the global reach and accessibility of online platforms have dramatically transformed the scale and complexity of child sexual exploitation, pushing the boundaries of conventional detection and prevention methods [4].

Traditional detection methods, such as victim reports and device-based investigations, are reactive and inadequate for the scale and complexity of online exploitation. The vast volume of data and evolving tactics of perpetrators require innovative and proactive solutions.

Generative AI, a rapidly advancing field of artificial intelligence, offers a promising solution in the fight against online child abuse. Artificial intelligence is increasingly recognized as a pivotal technology in this context, with applications ranging from the automated detection of predatory behavior using natural language processing to the detection of CSAM through image recognition systems [5]. Generative AI's ability to process large amounts of data, identify complex behavioral patterns, and generate new outputs makes it a valuable tool for detecting suspicious behavior and uncovering indicators of child sexual exploitation.

This technology holds significant promise in several key areas [6]:

- Automated CSAM Detection—AI algorithms can analyze images, videos, and text to flag both known and previously unseen instances of child sexual abuse material, including within encrypted environments.
- Online Predator Identification—By analyzing communication trends, linguistic markers, and digital behavior patterns, AI systems can help identify individuals engaged in grooming or exploitation attempts.
- Proactive Risk Assessment—Although still highly speculative and surrounded by profound ethical concerns, AI models could theoretically be used to identify behavioral markers associated with an increased risk of offending. However, any such application would require extensive ethical safeguards and governance to ensure that preventive actions are both responsible and justified.

Despite its potential, the use of generative AI in this highly sensitive area is fraught with challenges. Ethical issues related to privacy, algorithmic bias, the risk of erroneous identification, and the threat of misuse by malicious parties must be carefully navigated to ensure the technology is applied responsibly.

While the application of AI to combat online child exploitation is an active area of research, existing literature has predominantly focused on specific aspects of the problem. Previous studies have successfully applied traditional Machine Learning (ML) techniques, such as Support Vector Machines (SVMs) and rule-based systems, to detect predatory language in chat logs, often using established datasets like PAN2012 and Perverted Justice. More recent comprehensive surveys, such as the work by Borj et al. [7], have provided extensive overviews of these ML-based grooming detection techniques. Evidence assessments by researchers like Wolbers et al. have systematically reviewed the broader role of AI in identifying CSAM and aiding law enforcement. This body of work establishes the relevance and potential of AI in this critical domain.

However, a significant portion of this research predates the recent explosion of advanced Generative AI and Large Language Models (LLMs). While existing surveys have cataloged past methodologies, there is a clear gap in providing a forward-looking synthesis focused specifically on the unique capabilities and challenges presented by modern Gener-

ative AI. This paper aims to fill that gap by offering a distinct contribution that differs from the state-of-the-art in several key ways.

Our primary synthesis and novel perspective are consolidated in the later sections of this paper. Specifically, this survey distinguishes itself by:

- Identifying and analyzing the specific limitations of each individual work surveyed, presenting these findings systematically in our comparative analysis table.
- Providing a dedicated discussion of the collective technological and ethical risks associated with deploying Generative AI in such a sensitive context, including privacy concerns, algorithmic bias, and the potential for misuse.
- Proposing concrete and structured directions for future research, which are directly derived from the identified gaps and limitations in the current body of work.

To fully appreciate the advancements and capabilities of Generative AI in this domain, it is important to first understand the foundational approaches from which it evolved. Therefore, this review begins by examining the traditional research methods and classical Machine Learning models that have historically been applied to this problem. This foundational context provides a critical baseline for evaluating the paradigm shift introduced by Generative AI, particularly Large Language Models, which are the primary focus of the later sections of this analysis.

This paper explores the promise of generative AI in the fight against pedophilic crimes. We aim to provide an in-depth review of current research, methodologies, and challenges while exploring potential avenues for future development and ethical implementation. This study aims to address the following Research Question (RQ):

- RQ: How can generative AI be effectively used to detect and prevent pedophilia-related crimes in digital environments?

This work will critically analyze the existing literature and examine different methodologies and approaches. It will highlight both the challenges and future opportunities of using generative AI to protect children from online sexual exploitation. Used responsibly, AI has the potential to significantly enhance child protection in the digital age.

The remainder of this paper is organized as follows. Section 2 provides fundamental insights into the phenomenon of online grooming and the complexities of generative AI paradigms. Section 3 outlines the methodological framework used for a comprehensive review of the relevant literature. Section 4 presents the synthesized findings from the literature review. Sections 5 and 6 explore the advantages and inherent limitations of generative AI in this context. Section 7 distills the key takeaways drawn from the review. Section 8 outlines possible avenues for future research in this area. Finally, Section 9 summarizes the overarching conclusions drawn from the study. A list of abbreviations used throughout the paper is provided in the back matter, before the References.

2. Background

ML has been widely applied across various domains, including the detection of pedophilic crimes by analyzing textual, behavioral, and multimedia data. While traditional ML models, such as SVMs and decision trees, have shown promise in structured data contexts, the adoption of Generative AI for this purpose is a relatively novel and emerging direction. Recent research highlights the growing potential of Generative AI, particularly LLMs, in identifying grooming patterns, detecting abuse-related content, and generating synthetic data for training secure detection systems [8]. These advancements represent a significant shift in digital forensics capabilities and the ethical use of AI for child safety. This section provides an overview of the foundational concepts necessary to contextualize this study.

2.1. Online Grooming

Online grooming is a multifaceted and ever-changing process in which predators use deceptive methods to manipulate and exploit children for sexual purposes. The grooming process typically involves several stages [9]:

- Target Selection: Predators use online platforms and social media to identify vulnerable children;
- Relationship Building: Predators build trust with the child, often pretending to be a peer or offering emotional support;
- Risk Assessment: Predators assess the child's vulnerability and the level of parental supervision;
- Exclusivity: Predators attempt to isolate the child from their friends and family, fostering a sense of dependency;
- Sexualization: Predators introduce sexual topics into the conversation, gradually desensitizing the child to inappropriate content;
- Exploitation: The final stage involves attempts to meet the child in person or to coerce the child into producing explicit material.

Understanding these stages is critical to developing effective detection and intervention strategies.

2.2. From Traditional Machine Learning to Generative AI

To understand the impact of Generative AI, it is essential to first contextualize it against traditional approaches that have been historically used.

Traditional AI Models are mainly employed for data analysis, forecasting, and classification. These models use a variety of algorithms, including decision trees and SVMs, and can use supervised, unsupervised, or semi-supervised learning methods. Their primary applications are in data analysis across various domains like finance and healthcare. While they can perform well with medium-sized datasets, they face challenges that include overfitting and handling non-linear relationships [10]. Generative AI models represent a fascinating class of artificial intelligence algorithms that can create new and original content. Unlike conventional models, which focus on recognizing patterns in existing data, generative models can produce entirely new data similar to the examples on which they were trained. Their main purpose is data generation and synthesis for creative tasks, such as generating images, music, or text. These models are typically based on complex neural networks, like Variational Autoencoders (VAEs) or GPT, and may require larger amounts of data for effective learning [10].

The fundamental difference lies in their core purpose and capability. While traditional models focus on recognizing patterns to make predictions based on existing data, generative models create entirely new data. This is reflected in their primary focus: traditional models are built for data analysis, while generative models are built for data generation. However, each faces distinct challenges; generative models might struggle to capture complex relationships in tabular data, whereas traditional models are prone to issues like overfitting [10].

2.3. Large Language Models

An LLM is an advanced deep learning algorithm that is trained on large amounts of text data. This training gives LLMs the ability to [11]:

- Understand and generate human-like text: LLMs can participate in conversations, answer questions, summarize text, and even mimic different writing styles;
- Identify patterns and connections in language: This enables them to detect topics, sentiments, and potential warning signs in online exchanges;

- Learn and adapt: LLMs can be fine-tuned to specific datasets, such as conversations related to online grooming, to improve their accuracy in detecting suspicious behavior.
- While platforms like Perplexity AI, Google Gemini, and ChatGPT are primarily engineered for general-purpose information retrieval and dialogue generation, their underlying LLM architectures present valuable opportunities for specialized applications such as crime detection and child protection. Google Gemini, with its extensive context window and multimodal capabilities, enables the analysis of lengthy, nuanced conversations and supports the interpretation of diverse content formats—including text, images, and video—making it particularly effective in the detection of long-term grooming behavior and CSAM [6]. ChatGPT, known for its adaptability via fine-tuning, is well-suited for domain-specific implementations, such as detecting linguistic cues indicative of grooming or facilitating interactive simulation environments for proactive threat detection. Perplexity AI, leveraging real-time search integration and transparent source attribution, offers substantial potential for monitoring live online forums and chat spaces, where grooming often manifests. Collectively, these tools—when used with awareness of their distinct capabilities and limitations, can inform the development of more targeted and effective generative AI-driven safety systems.
- Regarding data retrieval, ChatGPT depends solely on its static training corpus, whereas both Gemini and Perplexity AI utilize dynamic search-based retrieval mechanisms. Gemini stands out with a significantly larger context window than the other two, allowing it to process and analyze complex documents in greater depth. Additionally, Gemini's multimodal processing sets it apart from the primarily text-based functionalities of ChatGPT and Perplexity. Cost and accessibility also vary: ChatGPT offers a freemium model, Perplexity remains free, and Gemini's pricing is tiered based on model version and access features.

It is important to understand how models with generative origins are used in detection tasks. Although these models, including LLMs, can generate new content, they are most effective in detection scenarios due to their sophisticated understanding of language, context, and patterns. In practice, a pre-trained generative model, such as BERT or Llama 2, is often used as a powerful feature extractor. Then, the model is fine-tuned on a specific dataset for a discriminative task, such as classifying a conversation as 'grooming' or 'non-grooming'. In this context, "Generative AI" does not necessarily refer to generating content for detection but rather to using these advanced generative pre-trained architectures for analytical and classification purposes.

2.4. Comparative Technical Analysis

LLMs are all fundamentally based on the revolutionary Transformer architecture [12], they exhibit significant differences in their specific designs, performance characteristics, and implementation challenges.

2.4.1. Architectural and Performance Differences

The primary architectural distinction between the models lies in their core components, training methodologies, and performance optimizations.

- ChatGPT is built upon the GPT architecture, which has evolved through several versions [12]. Its foundation consists of a series of transformer encoder layers, each utilizing a multi-head self-attention mechanism and a Feedforward Neural Network (FNN) [13]. The model is further aligned with human preferences using Reinforcement Learning from Human Feedback (RLHF) to enhance safety and produce more helpful responses [12,13].

- Gemini is also based on a decoder-only Transformer architecture [12]. Its design includes specific modifications for efficient training and inference on Tensor Processing Units (TPUs) and employs multi-query attention [13]. A key architectural feature is the integration of Retrieval-Augmented Generation (RAG), which grounds its responses in retrieved information to improve factual accuracy [12]. Gemini is designed as a natively multimodal system, capable of processing a combination of text, images, audio, and video [12]. In terms of performance, it is noted for prioritizing computational efficiency, potentially outperforming ChatGPT in speed and energy consumption [12].
- LLaMA, released by Meta AI, also uses the Transformer architecture but with several technical differentiators [13]. It employs Root Mean Square Layer Normalization (RMSQLN) instead of traditional layer normalization, and it uses the Swish-Gated Linear Unit (SwiGLU) activation function [13]. For positional information, it utilizes a Rotary Position Embedding (RoPE) scheme [13]. Regarding its context window, the original LLaMA was trained with a 2 K token context length, which was extended to 4 K tokens for LLaMA2 [13]. The practical implications of its computational requirements are significant; while compact models like LLaMA 7B can be run on local machines, more extensive versions demand impractical processing times, in the order of minutes per answer [13].

2.4.2. Implementation and Fine-Tuning Challenges

Despite their capabilities, the practical implementation of these LLMs reveals distinct challenges related to safety, bias, and resource requirements.

ChatGPT faces several known issues. The model can be vulnerable to "jailbreaking", where carefully crafted prompts are used to bypass its safety protocols and generate undesirable outputs [12]. It is also susceptible to perpetuating biases present in its training data and can sometimes struggle with logic and reasoning [12]. Furthermore, operating ChatGPT for certain applications can incur significant computational costs, potentially limiting its adoption [12].

The primary challenges associated with Gemini are related to its safety features and developmental stage. Its safety filters have been found to operate inconsistently across different languages and scenarios, with censorship being applied unpredictably, which raises concerns about its reliability [13]. As a more recent model, Gemini's limited exposure to extensive real-world data may hinder its ability to address nuanced biases when compared to more mature models [12].

LLaMA, on the other hand, exhibits a strong and consistent "optimistic bias" in its outputs, tending to rate scenarios positively regardless of their negative context [13]. A significant practical challenge is its hardware requirements; while compact versions can be run locally, the more extensive models demand impractically long periods to generate responses on personal computers, making them inaccessible to the average user [13].

The technical comparison in Table 1 is directly relevant to the context of detecting crimes related to pedophilia. It is crucial to understand the architectural differences, strengths, and limitations of leading LLMs—such as ChatGPT, Gemini, and LLaMA—to assess their suitability in sensitive domains like online grooming and CSAM detection. For example, Gemini's multimodal capabilities allow for the simultaneous analysis of image- and text-based content, which is essential for identifying disguised CSAM across platforms. Similarly, LLaMA's lightweight architecture may enable local deployment for privacy-sensitive scenarios, such as client-side filtering. Thus, this comparative overview informs model selection based not only on performance, but also on legal, ethical, and operational constraints in real-world deployments.

Table 1. Comparative Summary of LLM Technical Attributes and Challenges. This table provides a side-by-side comparison of the key technical features of the primary LLMs discussed in this review (ChatGPT, Gemini, and LLaMA). It covers their core architecture, key strengths, known weaknesses, and other relevant attributes.

Feature	ChatGPT (OpenAI)	Gemini (Google AI)	LLaMA (Meta AI)
Core Architecture	GPT with Reinforcement Learning from Human Feedback (RLHF) [12,13].	Transformer-based with Retrieval-Augmented Generation (RAG) and modifications for TPU efficiency [12,13].	Transformer-based with Root Mean Square Layer Normalization (RMSQLN) and Swish-Gated Linear Unit (SwiGLU) [13].
Multimodality	Primarily text-focused, with some multimodal capabilities noted as a relative weakness [12].	Natively multimodal, designed to process text, images, audio, and video [12].	Primarily text-focused, trained on text and code [13].
Key Strengths	Excels in conversational flow, creativity, and following instructions [12].	Prioritizes factual accuracy [12] and high computational efficiency (speed and energy) [12].	Shows high consistency in performance across different languages, suggesting easier transferability [13].
Key Weaknesses	Susceptible to "jailbreaking", logical inaccuracies, and high computational costs [12].	Exhibits inconsistent safety/censorship filters [13]. Being a more recent model, it has less real-world data exposure [12].	Exhibits a strong "optimistic bias". Larger models have impractical hardware requirements for local use [13].
Context Window	Able to handle prolonged interactions and maintain context [12].	Able to handle extended conversations and decipher intricate prompts [12].	Trained with a 2 K (LLaMA) or 4 K (LLaMA2) token context length [13].

3. Methodology

To address the research question: *How can generative AI be effectively used to detect and prevent pedophilia-related crimes in digital environments?* This study employed a systematic literature review approach. The goal was to map the current research applying AI techniques, particularly Generative AI and LLMs, to the detection of online grooming and related child exploitation crimes. This central objective guided the search strategy, selection criteria, and data characterization steps.

This review aimed to synthesize existing knowledge and critically evaluate the methodologies and results of relevant studies in order to understand the potential of generative AI in combating crimes related to pedophilia. Figure 1 illustrates the adopted methodology, which is detailed in the following subsections.

3.1. Data Source

Google Scholar (www.scholar.google.com) was selected as the primary data source for this review due to its extensive multidisciplinary academic literature, including peer-reviewed articles, conference papers, and technical reports relevant to AI applications and criminology.

In order to maintain the relevance and timeliness of the research, only studies published between 2010 and early 2025 were included, as this period reflects the significant advances in generative AI and machine learning applications.

In addition, the review focused exclusively on studies published in English, as this remains the dominant language in the scientific literature on artificial intelligence and criminology. Although this introduces a potential language bias, it was deemed necessary to maintain consistency in terminology and data interpretation.

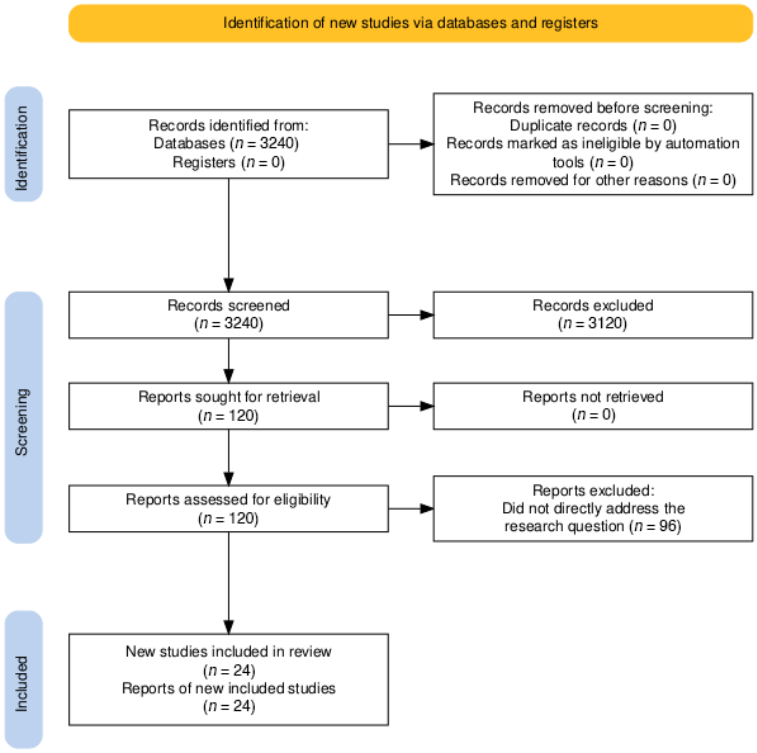

Figure 1. PRISMA 2020 flow diagram for the systematic literature review. The diagram illustrates the flow of information through the four phases of the review (Identification, Screening, Eligibility, and Included), detailing the number of records identified, screened, excluded, and ultimately included in the final analysis.

3.2. Search Query

The search strategy combined targeted keywords and Boolean operators to refine the results. Keywords such as "pedophile conversations", "AI", "data mining", "machine learning", and "pedophile crimes" were selected based on their direct relevance to the study's research question. A search in Google Scholar on 11 June 2025, with the query "AI AND ('Pedophile Conversations' OR 'Pedophile Crimes') AND ('Data Mining' OR 'Machine Learning')" returned 3240 results.

3.3. Inclusion Criteria

The following inclusion criteria were applied to ensure the relevance of the studies included in this review:

- Relevance to the topic: Only studies that directly addressed the use of AI to detect pedophilic crimes or analyze online conversations related to child sexual exploitation were included;
- Full-text availability: Studies were only included if their full text was available through Google Scholar or open-access repositories.

3.4. Exclusion Criteria

The following exclusion criteria were applied in order to narrow down the search results and to focus on relevant and high-quality research:

- Publication type: Materials such as books, book chapters, internal reports, theses/dissertations, citations, presentations, abstracts, and appendices were excluded from consideration;
- Language: Only studies published in English were included in the review.

3.5. Characterization of Selected Papers

The initial search in Google Scholar produced 3240 potentially relevant papers. After applying exclusion criteria, this number was reduced to 120. A subsequent screening process selected 24 papers for inclusion in the review.

This curated collection of studies represents the most relevant research on the use of AI to detect and combat pedophile crimes. Although relatively small in number, these studies provide insight into current advances in the field.

4. Literature Review: AI Approaches for Detection

This section provides a structured overview of previous studies on using AI to detect grooming behavior and online sexual exploitation. By organizing the literature into thematic subsections, we demonstrate how traditional machine learning and modern generative AI approaches have been applied in this field. Our goal is to contextualize the evolution of these techniques and identify areas where generative AI introduces novel capabilities or unresolved challenges.

This literature review compiles and analyzes scientific research on the application of AI in the detection of pedophilia. We have selected 19 papers to be examined, assessing the effectiveness, limitations, and ethical concerns associated with AI-based detection techniques. In this section, the papers are presented by subsections, and within each subsection, they are presented in chronological order based on their year of publication.

4.1. Identifying Grooming Behaviors and Risk Factors

This subsection examines research that focuses on the non-technical aspects of online grooming. The studies discussed here explore the identification of behavioral indicators exhibited by perpetrators and the assessment of risk factors associated with potential victims, including the development of specific measurement tools. Although it is not AI-based, this foundational work is intentionally included as it provides the essential groundwork for AI-driven detection systems. It effectively defines the specific 'red flag' behaviors and risk factors that advanced models, including Generative AI, aim to identify and automate the detection of. This serves as a critical baseline for the subsequent analysis.

Jeglic et al. [9] conducted a study to identify early indicators of sexual grooming by comparing the reported behaviors between adults who had experienced CSA and those who had not. The study included 913 participants, of whom 411 reported CSA and 502 did not. Those who reported CSA completed the SGS-V based on their personal experiences. In contrast, those who did not report CSA were randomly assigned to one of three categories: family member, non-family member, or community member. They were then asked to complete a modified version of the SGS-V about the adult male with whom they had the most frequent contact before the age of 18.

The study revealed a range of sexual grooming behaviors that distinguished the actions of adults involved in CSA from those who were not. The nature of the relationship between the adult and the child played a critical role in this distinction. One of the significant findings was the challenge of recognizing sexual grooming behaviors before the abuse takes place, as these actions often mirror everyday adult–child interactions. An important conclusion drawn from the study was that red flag grooming behaviors, particularly those aimed at desensitizing the child to physical touch and sexual content, can be identified in cases of CSA and are essential for prevention efforts.

Pasca et al. [14] conducted research to develop and validate the Online Grooming Risk (OGR) scale, a tool designed to detect online grooming in its earliest stages. The study focused on identifying and assessing risk factors associated with online grooming, such as low self-esteem and inadequate family support. The researchers also explored the

relationship between these risk factors and the likelihood of minors engaging in online sexual interactions with strangers.

The validation process involved a sample of 316 adolescents, with data analysis carried out using a non-parametric method known as Structural Equation Modeling based on Partial Least Squares (PLS-SEM). Key findings from the study revealed that the OGR scale is both a reliable and valid measure of online grooming risk. Furthermore, the research highlighted that family support is a protective factor, while low self-esteem emerged as a significant risk factor. The study emphasized the importance of identifying early warning signs of harmful behavior to prevent online grooming.

4.2. Machine Learning Models

This subsection presents a range of research that applies machine learning techniques, other than the deep learning approaches discussed next, to detect online grooming and related harmful online interactions. It includes studies using various algorithms such as rule-based systems, SVMs, and regression. Also included are comprehensive reviews of the field, proposals for integrated detection frameworks, and investigations into specialized system considerations such as privacy and human analyst support.

Kontostathis et al. [15] conducted a study to identify and categorize the tactics used by online sexual predators to establish relationships with minors. To achieve this, the researchers developed a rule-based system named ChatCoder, which was designed to classify and label lines of chat logs containing predatory language. Additionally, they experimented with machine learning algorithms, including decision trees and instance-based learning, to classify the posts.

The dataset for this study consisted of chat transcripts sourced from the Perverted Justice website. These transcripts contained conversations between convicted sexual offenders and volunteers posing as children. A random subset of 50 transcripts was selected, and 33 were used to create truth sets for training the machine learning models.

The ChatCoder 2 rule-based system achieved an average overall accuracy of 68.11% when compared to the hand-coded truth set. While the machine learning algorithms showed significant improvement when analyzing individual transcripts, they did not significantly outperform the rule-based ChatCoder system when the data was treated as a unified transcript. The results suggest that traditional machine learning approaches do not significantly improve the performance of the rule-based ChatCoder 2 system.

Upadhyay et al. [16] addressed the identification and mitigation of cyberbullying and online grooming through a complex application of machine learning and data mining methods. The cornerstone of their efforts is the conceptualization of an advanced "watchdog" application tailored to detect and thwart malicious online behavior. This application detects offensive language, inappropriate images, and unauthorized digital actions.

The research is grounded in three distinct detection paradigms: the Adult Image Detection Algorithm, the Irrelevant Posts Detection Algorithm, and a carefully crafted NLP Algorithm. Online grooming is defined as the act of an adult masquerading under a fabricated persona with the intent to manipulate and exploit a victim, often for predatory sexual purposes. To perform this classification, the study uses the Bad Words Dataset and the Sensitive Words Dataset, which serve as the foundational basis for categorizing user communications. Notably, the manuscript does not provide specific numerical results or the exact accuracies of the implemented algorithms. However, it does underscore the complexity of distinguishing legitimate content from illegitimate material and highlights the need for a cross-disciplinary approach to address these challenges.

Keyvanpour et al. [17] investigated automated techniques for detecting child abuse in online chat rooms, addressing a critical public safety concern. The goal was to create

a system capable of identifying predators and analyzing their network behavior. The research integrates text mining, pattern classification, and criminal psychology, using text preprocessing, feature extraction, and various classification algorithms.

The study examined detection models, including Suspicious Conversation Identification (SCI) and Victim From Predator Disclosure (VFP). Grooming behaviors were characterized using the Luring Communication Theory, which focuses on the processes of gaining access to victims, ensnaring them, and maintaining abusive relationships. The data used in the research came from the PAN-2012 competition. SVMs and Neural Networks provided the highest accuracy rates among the classification algorithms tested. Challenges to the study included the need for more nuanced linguistic analysis and the difficulty of handling unbalanced datasets. The researchers concluded that automated analysis of chat logs has significant potential for identifying online predators.

Ngejane et al. [18] focused on analyzing textual conversations, behavioral patterns, and sentiment analysis to uncover indicators of grooming behavior. The models demonstrated promising detection capabilities, achieving reported accuracy rates ranging from 73% to 98%. In addition, the research incorporated the Luring Communication Theory (LCT) model, which outlines five distinct phases that predators use to lure their victims, thereby enhancing the effectiveness of the detection algorithms.

While the findings highlight the potential of machine learning to identify online grooming behavior, the study also pointed out several challenges. These include the scarcity of labeled datasets, the critical need for practical feature engineering, and the difficulty of generalizing results to real-world scenarios. Overcoming these limitations is essential for improving the reliability and applicability of machine learning-based grooming detection systems.

Borj et al. [7] extensively explored techniques to detect online grooming. The paper analyzes chat logs and short texts between potential victims and predators. The researchers examine several machine learning algorithms. The paper also identifies and reviews several datasets used for grooming detection, and explores the importance of authorship profiling in the context of grooming detection, which involves examining individuals' unique writing styles and typing behaviors to uncover potential perpetrators.

The results show that machine learning models effectively identify online grooming with considerable accuracy. However, they also highlight several significant barriers to grooming detection, including data availability limitations, privacy concerns, unbalanced datasets, the non-uniform nature of chat logs, and the general unreliability of online data.

Zuo et al. [19] presented an innovative AI-powered framework designed to streamline the detection of online child grooming dialogues, addressing the inefficiencies and emotional toll associated with labor-intensive manual assessments. This system uses publicly available datasets for its training phase, and employs cutting-edge AI methods such as fuzzy-rough feature selection and fuzzy twin SVMs to scrutinize digital conversations. Its primary objective is to develop an automated solution that can detect grooming behavior in online interactions.

The study frames online child grooming as a complex process of emotional manipulation that ultimately facilitates sexual exploitation. The system seeks to identify these manipulative patterns by meticulously examining the text of conversations, a task that has traditionally been time-consuming and arduous. For the training process, the system used over 600 archived grooming chat logs sourced from the Perverted Justice website, as well as additional samples from the PAN13 dataset, resulting in a total of 1200 different documents. This research represents a significant leap forward in the use of AI in digital forensics, offering a potential solution for identifying online grooming behavior while highlighting the continued need for innovation to effectively address this pressing issue.

Pranoto et al. [20] introduced a logistic regression approach to classify online dialogues as either grooming or non-grooming. The research involved the analysis of 160 online transcripts of conversations randomly obtained from http://www.perverted-justice.com (accessed on 6 May 2025) and www.literotika.com (accessed on 6 May 2025), which were categorized based on 20 different grooming characteristics and divided into training and test datasets. The study identified five key grooming features using a stepwise regression combined with three methods and paired t-tests. The model demonstrated impressive accuracy, reaching 95%, underscoring its potential to detect online grooming conversations. However, the study faced challenges, including the lack of child-specific language within grooming exchanges and the overlap of grooming features within non-grooming dialogues.

Fauzi et al. [21] introduced a two-stage methodology that uses a SVM and a feature ensemble strategy to identify predators in online chat interactions. The framework operates in two phases: the first focuses on detecting predatory conversations, while the second differentiates between predators and their victims.

The study integrates lexical and behavioral features, with the BoW representation emerging as the most compelling feature set. The SVM was chosen because of its proven effectiveness in addressing text classification challenges. The PAN12 Sexual Predator Identification dataset, initially developed for a competition at the 2012 Conference and Labs of the Evaluation Forum (CLEF), served as the basis for the experiments. The results show that the BoW approach, combined with tf-idf term weighting, provides superior performance for the PCI and VPD tasks. Specifically, the method achieved F0.5 scores of 0.9893 for PCI and 0.9798 for VPD, underscoring its high accuracy and reliability. Although the feature ensemble technique shows significant potential, the study highlights the importance of mitigating the impact of noisy features to optimize performance.

Ebrahimi et al. [22] introduced a semi-supervised anomaly detection framework to identify predatory chat documents. This framework effectively addresses the limitations of supervised techniques that rely on extensive labeled data for predatory and non-predatory categories. Their approach uses a one-class SVM trained exclusively on predatory chat data, thereby eliminating the need to label non-predatory conversations.

The researchers evaluated their methodology using the PAN-2012 dataset, a well-known dataset curated for sexual predator identification tasks. The data was preprocessed using RapidMiner, while the SVM model was implemented using LibSVM. The experimental results showed that the one-class SVM outperformed the performance of the baseline Naive Bayes classifier and achieved results comparable to those of the binary SVM classifier, reaching an F-measure of 75% after incorporating a noise removal step. This study underscores the practicality of semi-supervised anomaly detection for identifying predatory chat documents, highlighting its ability to perform effectively with significantly less labeled data.

Nasir et al. [23] proposed a method for classifying and identifying sexually exploitative content within textual exchanges on these platforms. The primary goal of this study was to develop an algorithm that can accurately detect grooming content while also assessing the performance of the classification technique.

The detection models employed include n-gram and sequence models to enhance the classification precision. The data was gathered from several widely used social media platforms, including YouTube, Instagram, and Twitter. The results show that the n-gram model achieved an accuracy rate of 81.6%, while the sequence model had a slightly lower accuracy of 69.8%. The study acknowledges the challenges of detecting grooming behavior due to the sheer volume of online content shared daily, which complicates the identification process. It demonstrates how raw social media datasets can be used to develop and apply classification algorithms.

Chehbouni et al. [24] proposed a privacy-preserving, cross-device federated learning framework designed for the early detection of sexual predators. The core objective of the study was to develop a decentralized, privacy-preserving method using FL to detect sexual predators in online communications at an early stage. It seeks to balance the need for protecting children from cyber threats with the paramount importance of preserving user privacy.

Federated learning is utilized to train a context-aware language model in a distributed manner. The focus of the study is on identifying linguistic markers within online dialogues that could indicate grooming behavior. By detecting these early signs, the system aims to facilitate timely intervention. The study demonstrates that federated learning offers a promising approach for the early detection of sexual predators, providing a balance between detection accuracy and user privacy. The proposed framework performs comparably to centralized models while ensuring that user data is protected.

Eilifsen et al. [25] conducted an investigation into the effectiveness of machine learning algorithms for the early detection of cyberbullying in digital dialogues. Employing a sliding window methodology, the research examined segments of fixed-length message windows, incrementally traversing through conversational threads. This analytical approach facilitated the dissection of message clusters to identify suspicious patterns with temporal precision.

The PAN2012 dataset served as the empirical basis for the study, providing a repository of chat dialogues that distinguish "normal" interactions from those characterized by potential predatory behavior, particularly involving sexual predators. Among the algorithms tested, the Multinomial Naive Bayes with Term Frequency (mult-nb-tf) model emerged as the most proficient, demonstrating superior accuracy over sliding window lengths of 5 and 10 messages. Despite the promising results, the study acknowledged several limitations. The most significant of these were the modest sample size of conversations and the rigidity of the fixed sliding window dimensions.

4.3. Deep Learning and Large Language Models

This final subsection focuses specifically on research using deep learning architectures, such as Long Short-Term Memory networks (LSTMs) and transformer models such as BERT. It also covers the evaluation and application of LLMs to improve the detection of online grooming behaviors.

Liu et al. [26] presented an innovative SPI technique to combat the escalating threat of cybercriminal behavior on social media platforms. Their method integrated sentiment analysis, a novel strategy for generating sentence vectors, and LSTM-RNNs. In this approach, an LSTM-RNN language model generates sentence vectors, which are then utilized by another LSTM-RNN classifier to detect potentially harmful conversations. Fasttext is used to perform sentiment analysis on the conversation content, generating sentiment scores that contribute to the identification of possible predators. The proposed approach exhibits remarkable accuracy and precision, outperforming previous results in the SPI competition. The study utilized the PAN2012 dataset, as well as the IMDB movie review dataset, for training.

The study yielded significant results, including a perplexity score of 10.948 for the language model, an impressive 99.43% accuracy in detecting conversations, and effective identification of sexual predators. However, challenges such as noise in the dataset and the inherent complexity of the task remain. The study concludes that the combined approach offers an effective solution for identifying sexual predators in online spaces.

Puentes et al. [6] explored the potential of automated systems to support human analysts in managing the troubling content associated with crimes such as sextortion, sexting, grooming, and sexual cyberbullying. Recognizing the emotional burden and time-consuming nature of manual analysis, the authors proposed an innovative tool powered by a BERT-based LLM to categorize abuse reports based on subject matter, criminal severity, and harm. This model was trained and evaluated using 1196 real reports sourced from "Te Protejo", a Colombian hotline dedicated to combating online child sexual abuse.

The study provides a detailed account of the data annotation process, highlighting the critical role of a multidisciplinary team in ensuring both accuracy and sensitivity. To overcome challenges such as data sensitivity and class imbalance, the authors used a data augmentation method, specifically, randomly deleting words from the complaints to create a more diverse set of training examples, thereby increasing the robustness of the model. The results show that fine-tuning the LLM improves classification accuracy and reduces false positive rates, suggesting the potential for such automated tools to make a meaningful contribution to combating online child sexual abuse.

Prosser and Edwards [11] investigated the effectiveness of large LLMs in combating online grooming by assessing their ability to detect such behavior and provide guidance. They evaluated six LLMs, including ChatGPT, PaLM 2, and Claude 2, focusing on how the design of the prompts affected the models' ability to detect grooming and provide appropriate advice to children. The researchers tested the models with 100 prompts based on actual conversations between adults posing as minors and online predators.

The results showed that none of the models adequately prevented online grooming. Specifically, closed-source models were found to be overly cautious, while open-source models had a greater risk of generating harmful responses. A major challenge identified was the inconsistency of responses from the LLMs, particularly when handling conversations of moderate to low risk. This variability, even within the same model, underscores the need for ongoing research to refine the accuracy and safety of LLMs in addressing online grooming situations.

Vogt et al. [27] focused on the early detection of sexual predators in online chats to protect children from digital grooming. The authors aimed to develop a system capable of analyzing ongoing chat conversations in real-time, detecting grooming attempts, and intervening before these interactions escalate into sexual abuse. To train and evaluate their system, the researchers developed a novel dataset called PANC, which merged data from two existing datasets: PAN12 and ChatCoder2.

The system uses a BERT model to examine sliding windows of chat messages and classifies each window as predatory or non-predatory. Among the models tested, the BERT-base model showed superior performance in early detection. The BERT-base and BERT-large models outperformed current state-of-the-art methods in non-early detection scenarios. The effectiveness of the models was evaluated using the F1 score, precision, recall, and processing speed. Despite challenges such as obtaining authentic grooming chat data and the lack of fully representative negative chats in the dataset, the research demonstrated the viability of using BERT models for early detection of sexual predators.

Aarnseth et al. [28] investigated the fine-tuning of the BERT language model for detecting cyber grooming, with a specific focus on how informal language impacts its performance. The primary objective was to assess whether BERT could be trained to understand abbreviations and slang within predatory online chats and whether replacing such language with formal text would improve detection rates. Using the PyTorch framework, various models including BERT-base, RoBERTa, and ALBERT were fine-tuned on the PAN12 and AIBA AS datasets, which contain both formal and highly informal

conversations. Methodologies also included data augmentation through techniques like back-translation to address significant data imbalances.

The study evaluated model performance using precision, recall, and F1-score as its primary metrics for accuracy. The BERT-base model demonstrated strong performance, achieving a top F1-score of 0.86 when trained on the full PAN12 dataset. A critical finding was that its performance remained comparable on subsets of the formal PAN12 data (0.63 F1-score) and the highly informal AIBA AS data (0.62 F1-score), indicating a high tolerance for slang. The research concluded that replacing abbreviations degraded performance and that BERT's architecture provides a robust understanding of language without extensive pre-processing. Key challenges included data imbalance and the variability of slang. Future work should focus on methods to add extra weight to high-risk keywords and create more advanced systems for translating informal language.

Borj et al. [29] introduced a contrastive learning framework for sentence-based feature extraction to identify predatory online conversations, effectively handling misspellings. Its core methodology utilized RoBERTa encoders and supervised SimCSE to train an SVM classifier, achieving 0.99 accuracy and a 0.96 $F_{0.5}$ Score.

The study aimed to develop a robust grooming detection system using the PAN 2012 dataset, with feature extraction performed via SimCSE (utilizing RoBERTa/BERT) and classification achieved through SVM and other models. Fusion techniques, like the sum rule, further boosted performance to 0.99 accuracy and a 0.98 $F_{0.5}$-score. The research highlights the efficacy of the RoBERTa-SimCSE-SVM setup and ensemble fusion, emphasizing the $F_{0.5}$-score's law enforcement relevance. It also notes challenges like context dependency and data imbalance, concluding that sentence-level embeddings are superior.

Nguyen et al. [30] investigated the detection of online sexual predatory chats and abusive language using the Llama 2 large language model. Their approach focused on fine-tuning the pre-trained Llama 2 7B-parameter model with different datasets that varied in size, language, and degree of imbalance. Their study aimed to overcome the limitations of traditional methods, such as reliance on manual feature extraction and classifier design, and to address the lack of a universal model that can consistently perform across different datasets and languages.

The study employed three distinct datasets: the PAN 2012 dataset for detecting sexual predatory chats and the Roman Urdu and Urdu datasets for identifying abusive language. The fine-tuned Llama 2 model delivered exceptional performance on all three datasets, surpassing the results of existing state-of-the-art methods. The study highlighted the potential of this approach for detecting online sexual predators and offensive language in both English and non-English languages.

Rho et al. [31] introduced a framework to combat online grooming on social networking services by integrating OCR with a deep learning-based NLP model, KcELECTRA, specifically tailored for the Korean language. The objective was to accurately detect grooming conversations, which are characterized by a six-stage process of manipulation, to protect vulnerable users. The methodology involves extracting text from SNS chat images using OCR and then classifying the content using the fine-tuned KcELECTRA model. The model was trained on extensive Korean datasets from AI-Hub, including SNS conversations and hate speech. The proposed KcELECTRA model was benchmarked against several others, including RoBERTa, BLOOMz, and DeBERTa-v3. Experimental results showed that the framework achieved a high accuracy of 0.953, outperforming existing transformer-based models. Despite this success, the study faced challenges such as the limited availability of Korean-specific grooming data and the insufficient accuracy of OCR technology, which can misinterpret nuanced expressions. Key conclusions highlight the framework's competitive efficiency and its potential to create safer online environments by focusing on linguistic cues

while upholding user privacy. Future work will focus on enhancing multimodal detection, addressing data privacy through synthetic data generation, and improving context-aware classification by analyzing longer conversational sequences.

Yang et al. [32] aimed to develop and optimize algorithms for the early detection of online grooming in Korean conversations, using the PAN12 Korean dataset. Researchers evaluated seven BERT-based models and three LLMs, comparing a conventional "window method" against a novel "memory method". The window method references a fixed number of consecutive previous sentences, while the memory method retrieves sentences based on semantic similarity. This approach was designed to enhance accuracy and speed by focusing on relevant context. To address memory constraints during training, the Quantized Low-Rank Adaptation (QLoRA) method was applied to the LLMs,

The results demonstrated that the optimal method depends on the model's architecture. For instance, with a window size of 10, the Roberta-large model achieved an F_1 score of 0.85, accuracy of 0.81, and speed of 0.68. In contrast, the memory method with the same parameters resulted in an F_1 score of 0.84, accuracy of 0.79, and speed of 0.71. The study also introduced two key evaluation metrics: the latency-weighted F_1 ($F_{latency}$), which combines F_1 score and detection speed, and the Human-to-Model Ratio (HMR), which compares model detection speed to human judgment. A key conclusion was that the memory method is more suitable for BERT-based models in long conversations, while the window method is highly effective for LLMs due to their larger token capacity.

Hamm et al. [33] advanced the crucial field of online grooming detection by systematically comparing the performance of traditional machine learning with modern LLMs, specifically focusing on how predator tone influences detection accuracy. Utilizing the PAN12 chat log dataset, the study employed a SVM and Meta's LLaMA 3.2 1B large language model, after sorting conversations into positive and negative sentiments with a DistilBERT classifier. The results conclusively show that the LLaMA model significantly outperforms the SVM, achieving an F_1 score of 0.99 compared to the SVM's 0.94, demonstrating deep learning's superior ability to capture the complex linguistic patterns of grooming.

A primary contribution is the discovery that predators employ distinct strategies based on tone, with negative-toned chats containing more nuanced and harder-to-detect patterns. This finding suggests that positive and negative-toned approaches are structurally dissimilar. The LLaMA 3.2 1B model established a new performance benchmark, outperforming previous models with an F_1 score of 0.99 in author detection. The study concluded that LLMs offer a more effective path for identifying grooming behavior and revealed that diverse predator strategies exist. Future work should focus on validating these findings on contemporary datasets and exploring more granular emotional classifiers.

4.4. Comparative Analysis

Table 2 provides a comparative overview of the key studies examined in this review, highlighting the AI methods employed, datasets used, reported accuracy rates, and identified strengths and limitations. This comparative analysis helps to understand the evolution of AI techniques applied to the detection of pedophilia crimes and highlights existing gaps in current research.

To provide a clearer overview of the landscape, the following Table 3 maps the AI techniques discussed in this survey to the characteristics of the datasets used for their training and evaluation. It highlights the domain, language, and size of the datasets where such information is available in the reviewed literature.

Table 2. Comparative Analysis of Key Artificial Intelligence Studies for Pedophilia Crimes Detection. This table summarizes the 24 studies included in this review, detailing the AI method employed, the dataset used, the reported accuracy or key performance metric, and the primary strengths and limitations identified for each study.

Study	AI Method	Dataset Used	Accuracy (%)	Key Strength	Limitation
Jeglic et al. (2022) [9]	SGS-V Scale	CSA Victim Reports	88	Early grooming indicators	Recall bias
Pasca et al. (2022) [14]	OGR Scale	Survey-Based	89	Focus on risk factors	Limited generalizability
Kontostathis et al. (2011) [15]	Rule-based system ChatCoder, decision trees, instance-based learning	Perverted Justice	68	ChatCoder 2's classifications were more accurate than the hand-coded truth set	ML algorithms did not significantly outperform the rule-based ChatCoder system
Upadhyay et al. (2017) [16]	Machine learning, image/content filtering, NLP	Bad Words Dataset, Sensitive Words Dataset	Not provided	Promising potential to protect youth from online grooming and cyberbullying	Complexity of distinguishing lawful content from illicit material
Keyvanpour et al. (2018) [17]	Text Mining and SVM	PAN2012	85	Network behavior analysis	Imbalanced datasets
Ngejane et al. (2018) [18]	SVMs, k-NNs, CNNs, semi-supervised anomaly detection	Perverted Justice, PAN2012 and MovieStarPlanet	98	Models demonstrated promising detection capabilities	Scarcity of labeled datasets, and the difficulty of generalizing results to real scenarios
Borj et al. (2019) [7]	Neural Networks	Perverted Justice and PAN2012	90	Diverse algorithm use	Privacy concerns
Zuo et al. (2019) [19]	Fuzzy-rough feature selection and fuzzy twin SVMs	Perverted Justice, PAN13-Author-Profiling	61	AI technologies integration promise for overcoming the limitations of traditional detection methods	The ever-evolving nature of grooming language and chat formats; Need for a larger representative dataset
Pranoto et al. (2020) [20]	Logistic Regression	Perverted Justice	95	High detection accuracy	Lack of child-specific language
Fauzi et al. (2020) [21]	SVM with BoW	PAN2012	98	High accuracy in VPD	Noisy data impact
Ebrahimi et al. (2021) [22]	One-Class SVM	PAN2012	75	Works with less labeled data	Lower F1-score
Nasir et al. (2022) [23]	Text mining, text classification, neural network, n-gram and sequence models	YouTube, Instagram, and Twitter	82	Enhanced precision with n-gram and sequence models	Sequence model accuracy lower than n-gram model
Chebouni et al. (2022) [24]	Logistic Regression	PANC	79	Privacy-preserving approach, addresses non-IID data, early detection focus	Limitations of PANC, biases in BERT, computational cost of FL
Eilifsen et al. (2023) [25]	Naive Bayes, Tree-Based Models, Neural Networks, and Dynamic Trust Model	PAN2012	88	Multinomial Naive Bayes model with Term Frequency (mult-nb-tf) demonstrated superior accuracy	Modest sample size, rigidity of fixed sliding window dimensions
Liu et al. (2017) [26]	LSTM-RNN for sentence vectors and conversation classification; Fast text for sentiment analysis	PAN2012 dataset; IMDB movie review dataset (for sentence vector evaluation)	99.43 (chat); 98.35 (Predator); 83.2 (IMDB)	Effectively captures dependencies in conversations; Sentence vectors reduce input dimensionality; Sentiment analysis enhances predator identity	Performance variability on IMDB dataset; Complexity in processing noisy and varied online chat data
Puentes et al. (2021) [6]	BERT-based LLM	Te Protejo Dataset	93	Handles sensitive data	Data sensitivity issues
Vogt et al. (2021) [27]	BERT Transformer	PAN2012 and ChatCoder2	92	Early-stage detection	Limited real-world data
Aarnseth et al. (2023) [28]	BERT	PAN2012 and AIBA AS	Uses F1-score; F1-score of 86	Model is highly robust, proving effective at detecting cyber grooming even when conversations contain a large amount of informal language and slang	Data imbalance
Borj et al. (2023) [29]	RoBERTa/BERT, SVM	PAN2012	99	Robust sentence-based feature extraction for high true-positive predatory conversation detection.	Accurately interpreting nuanced online conversations is challenging due to the context-dependent nature of language.
Nguyen et al. (2023) [30]	Llama 2 LLM	PAN2012 and Urdu	96	Multilingual capabilities	Data imbalance in non-English sets
Prosser and Edwards (2024) [11]	ChatGPT3.5, 4, PaLM2, Claude2, LLaMA2, Mistral	ChatCoder2 and Perverted Justice	Not provided	Prompt Design Impact Analysis	No LLM reliable; Consistency/Safety Issues

Table 2. *Cont.*

Study	AI Method	Dataset Used	Accuracy (%)	Key Strength	Limitation
Rho et al. (2025) [31]	KcELECTRA	A combination of Korean datasets	95	The framework outperforms existing transformer-based models in accuracy and is specifically optimized for the nuances of colloquial Korean language found in SNS conversations	Data scarcity and semantic similarities between hate speech and sexually explicit content pose challenges
Yang et al. (2025) [32]	BERT and LLM models	PAN2012 dataset translated into Korean augmented with Korean SNS	81	Proposes a new metric, Human-to-Model Ratio (HMR), for a more nuanced evaluation of detection speed	Models struggle to interpret figurative language used in grooming conversations
Hamm et al., (2025) [33]	LLaMA 3.2 1B	PAN2012	99	The LLaMA 3.2 1B model outperforms both traditional machine learning models and previous, larger LLMs in grooming detection	The research relies on the PAN12 dataset, which is over a decade old and features law enforcement officers posing as victims, not real children

Table 3. AI Techniques versus Dataset Characteristics. This table maps the AI techniques from the surveyed literature to the specific characteristics of the datasets used for their evaluation, including the domain (e.g., online chat logs, social media), dataset language, and reported size.

Study	AI Method	Dataset Used	Dataset Domain	Dataset Language	Size
Jeglic et al. (2022) [9]	SGS-V Scale	CSA Victim Reports	Victim survey reports	English	913 participants
Pasca et al. (2022) [14]	OGR Scale	Survey-Based	Adolescent survey	Not Specified	316 adolescents
Kontostathis et al. (2011) [15]	Rule-based system ChatCoder, decision trees, instance-based learning	Perverted Justice	Online chat logs	English	50 transcripts selected (33 used)
Upadhyay et al. (2017) [16]	Machine learning, image/content filtering, NLP	Bad Words Dataset, Sensitive Words Dataset	Social media content	Not Specified	Not Specified
Keyvanpour et al. (2018) [17]	Text Mining and SVM	PAN2012	Online chat logs	English	288,142 lines (PAN2012 training set)
Ngejane et al. (2018) [18]	SVMs, k-NNs, CNNs, semi-supervised anomaly detection	Perverted Justice, PAN2012 and MovieStarPlanet	Online chat logs, social gaming	English	288,142 lines (PAN2012 training set)
Borj et al. (2019) [7]	Neural Networks	Perverted Justice and PAN2012	Online chat logs	English	288,142 lines (PAN2012 training set)
Zuo et al. (2019) [19]	Fuzzy-rough feature selection and fuzzy twin SVMs	Perverted Justice, PAN13-Author-Profiling	Online chat logs	English	1200 documents
Pranoto et al. (2020) [20]	Logistic Regression	Perverted Justice	Online chat/story logs	English	160 transcripts
Fauzi et al. (2020) [21]	SVM with BoW	PAN2012	Online chat logs	English	288,142 lines (PAN2012 training set)
Ebrahimi et al. (2021) [22]	One-Class SVM	PAN2012	Online chat logs	English	288,142 lines (PAN2012 training set)
Nasir et al. (2022) [23]	Text mining, text classification, neural network, n-gram and sequence models	YouTube, Instagram, and Twitter	Social media comments	Not Specified	1000 comments
Chebouni et al. (2022) [24]	Logistic Regression	PANC	Online chat logs	English	32,510 segments
Eilifsen et al. (2023) [25]	Naive Bayes, Tree-Based Models, Neural Networks, and Dynamic Trust Model	PAN2012	Online chat logs	English	288,142 lines (PAN2012 training set)
Liu et al. (2017) [26]	LSTM-RNN for sentence vectors and conversation classification; Fast text for sentiment analysis	PAN2012 dataset; IMDB movie review dataset (for sentence vector evaluation)	Online chat logs, movie reviews	English	PAN2012 (288,142 lines); IMDB (50,000 reviews)
Puentes et al. (2021) [6]	BERT-based LLM	Te Protejo Dataset	Child abuse hotline reports	Spanish	1196 reports
Vogt et al. (2021) [27]	BERT Transformer	PAN2012 and ChatCoder2	Online chat logs	English	32,510 segments
Aarnseth et al. (2023) [28]	BERT	PAN2012 and AIBA AS	Online chat logs	English	PAN12 (288,142 lines); AIBA AS (4429 messages)

Table 3. *Cont.*

Study	AI Method	Dataset Used	Dataset Domain	Dataset Language	Size
Borj et al. (2023) [29]	RoBERTa/BERT, SVM	PAN2012	Online chat logs	English	288,142 lines (PAN2012 training set)
Nguyen et al. (2023) [30]	Llama 2 LLM	PAN2012 and Urdu	Online chat/abusive texts	English, Urdu	PAN2012 (288,142 lines); Roman Urdu (20,000 samples); Urdu (15,000 samples)
Prosser and Edwards (2024) [11]	ChatGPT3.5, 4, PaLM2, Claude2, LLaMA2, Mistral	ChatCoder2 and Perverted Justice	Online chat logs	English	100 prompts
Rho et al. (2025) [31]	KcELECTRA	A combination of Korean datasets	Social media chat images	Korean	~1.7 million sentences + 9400 comments
Yang et al. (2025) [32]	BERT and LLM models	PAN2012 dataset translated into Korean augmented with Korean SNS	Online chat logs	Korean	57 training + 147 test convos (+888 SNS convos)
Hamm et al., (2025) [33]	LLaMA 3.2 1B	PAN2012	Online chat logs	English	288,142 lines (PAN2012 training set)

The limitations identified in the surveyed literature can be categorized into several key areas. The following Table 4 organizes these challenges into technical, data-related, ethical/privacy, and implementation categories, providing a structured view of the hurdles facing the field.

Table 4. Categorization of Identified Limitations. This table organizes the challenges and limitations reported in the surveyed studies into four distinct categories: Data-Related, Technical/Methodological, Ethical and Privacy, and Implementation and Generalization, providing example studies for each described limitation.

Category	Limitation	Studies Citing This Limitation
Data-Related	Data Scarcity, Quality and Imbalance: Lack of large, high-quality, labeled, and balanced datasets.	Keyvanpour et al. (2018) [17], Ngejane et al. [18], Zuo et al. [19], Aarnseth et al. [28], Nguyen et al. [30], Rho et al. [31]
	Outdated Datasets: Reliance on older datasets (e.g., PAN2012) that may not reflect current online behaviors.	Hamm et al. [33]
	Lack of Realism: Datasets using volunteers instead of real victims can lack authenticity.	Vogt et al. [27], Hamm et al. [33]
	Noisy and Unreliable Data: User-generated content is often unstructured, contains misspellings, and slang.	Liu et al. [26], Fauzi et al. [21]
Technical/Methodological	Model Reliability and Consistency: LLMs can provide inconsistent or overly cautious responses and are prone to "hallucination".	Prosser and Edwards [11]
	Contextual Understanding: Difficulty in interpreting nuanced, figurative, or evolving language.	Borj et al. [7], Borj et al. [29], Yang et al. [32]
	Difficulty of the Task: The inherent complexity of distinguishing legitimate content from harmful content.	Upadhyay et al. [16], Liu et al. [26]
	Recall Bias and Methodological Rigidity: Limitations in non-AI studies, such as recall bias in surveys or rigid window sizes.	Jeglic et al. [9], Eilifsen et al. [25]
Ethical and Privacy	Privacy Concerns: Monitoring online communications raises significant privacy issues.	Borj et al. [7]
	Algorithmic Bias: Models can have inherent biases (e.g., from BERT) that need mitigation.	Chehbouni et al. [24]
	Data Sensitivity: Handling real reports from abuse hotlines requires extreme care and restricts data availability.	Puentes et al. [6]
Implementation and Generalization	Limited Generalizability: Difficulty in applying models trained on one dataset to different, real-world scenarios.	Pasca et al. [14], Ngejane et al. [18]
	Computational Cost: The cost of training and deploying large or federated models can be a significant barrier.	Chehbouni et al. [24]

The surveyed AI-based methods can be classified by their intervention timeline, which determines whether they act retrospectively on existing data or proactively to prevent future harm. This distinction is crucial for understanding their practical application by law enforcement and child protection agencies (Table 5).

Table 5. Classification of Intervention Timelines. This table classifies the surveyed studies based on their primary intervention timeline, distinguishing between foundational risk factor identification, retrospective analysis of historical data, and real-time or early detection systems.

Intervention Timeline	Description	Approach/Key Feature	Studies
Foundational/Risk Factor Identification	Research focused on identifying grooming behaviors and risk factors without creating an automated detection system. This work is foundational for later AI models.	Development and validation of scales (SGS-V, OGR) through surveys and statistical analysis.	Jeglic et al. [9], Pasca et al. [14]
Retrospective Analysis	Analysis of static, historical datasets of conversations or content. Useful for investigations, pattern discovery, and benchmarking models.	Classification and analysis of established datasets like PAN2012, Perverted Justice, or social media data.	Keyvanpour et al. [17], Fauzi et al. [21], Ebrahimi et al. [22], Liu et al. [26], Borj et al. [29], Hamm et al. [33]
Real-Time/Early Detection	Monitoring and analyzing conversations as they occur to flag suspicious activity immediately, enabling intervention before harm escalates.	Use of sliding window techniques, federated learning on user devices, or memory-based context retrieval.	Chehbouni et al. [24], Eilifsen et al. [25], Vogt et al. [27], Yang et al. [32]
Multilingual and Multimodal Detection	Systems designed specifically to handle different languages or content types beyond text (e.g., images).	Fine-tuning language-specific models (e.g., Korean) or using OCR to extract text from images.	Nguyen et al. [30], Rho et al. [31]
LLM Capability Assessment	Studies focused not on building a detector, but on evaluating the inherent capabilities and safety of existing, general-purpose LLMs for this task.	Prompt-based testing of commercial LLMs like ChatGPT, PaLM2, etc.	Prosser and Edwards [11]

After exploring the state of the art in traditional ML and emerging GenAI techniques for detecting grooming behavior, the following section will highlight the benefits that these new models bring to this complex domain.

5. Benefits of Generative AI in Detecting Pedophilia

This section, along with the one that follows, explores the research question RQ1: "How can generative AI be effectively used to detect and prevent pedophilia-related crimes in digital environments?"

From a law enforcement and child protection perspective, many potential benefits of generative AI have been recognized, which we detail in the following subsections.

5.1. Automating CSAM Detection

The advanced architectures that underpin Generative AI, such as Transformers, are revolutionizing the automated detection of CSAM. For textual content, LLMs can directly analyze discussions for harmful patterns. While detection of visual and audio content is fundamentally a discriminative task, state-of-the-art classification models increasingly leverage the same robust architectures that originated in the generative domain. This technological capability outperforms traditional methods, offering superior speed and comprehensive coverage by facilitating the rapid identification of illicit material. For example, the work of Puentes et al. [6] presents a BERT-based LLM that was trained using real abuse reports from the 'Te Protejo' hotline in Colombia. This model successfully classifies complaint types, severity, and urgency. This example illustrates how generative

models can be fine-tuned to assist with real-world investigations, thereby reducing the workload and emotional burden on analysts.

Innovative approaches are exploring the use of metadata, such as file paths, to train machine learning models for CSAM detection, which could offer advantages in scenarios with strict legal and ethical limitations on accessing direct CSAM imagery [34]. Such methods aim to maintain high accuracy while minimizing the need to handle sensitive visual content directly during model development. The proliferation of AI-generated CSAM (AIG-CSAM) itself presents new vectors of harm. It complicates detection efforts, necessitating a review of both technological and policy-based solutions to combat this evolving threat [35].

While batch processing techniques are effective for retrospective analysis of large datasets, they often fail to detect CSAM content in real-time, which limits their preventive capabilities. In contrast, real-time AI models, such as those using streaming data architectures (e.g., Apache Kafka integrated with transformer models), enable continuous monitoring and immediate flagging of suspicious activity. Although real-time systems require more computing resources and have higher false positive rates, they offer a critical advantage in preventing live-streamed abuse and enabling faster law enforcement responses. The choice between batch and real-time processing should, therefore, be based on the specific operational requirements and available infrastructure of the implementing agency.

5.2. Spotting Online Predators

Generative AI offers a promising solution for identifying and detecting online predators by leveraging its ability to analyze large datasets, recognize intricate patterns, and generate novel content, which is primarily useful for creating realistic synthetic training data, such as simulated grooming conversations, to augment sparse or imbalanced datasets and build more robust detection models. As an example, Borj et al. [29] implemented a SimCSE-based contrastive embedding model with RoBERTa encoders—a powerful architecture from the generative AI domain—to extract sentence-level features from grooming conversations. They used these features to train an SVM classifier that achieved 99% accuracy in identifying predatory chats. This demonstrates how LLMs can be repurposed as reliable detectors.

By analyzing communication trends, language use, and online activity, AI can help identify individuals involved in grooming activities or seeking to exploit children for sexual purposes.

This proactive approach enables timely intervention, providing an opportunity to protect children from harm before it occurs. In addition to identifying existing cases of exploitation, generative AI also plays a role in uncovering new tactics used by predators, ultimately increasing the effectiveness of child protection strategies [6].

5.3. Predicting Potential Offenders

Generative AI has the potential to analyze online signals and identify high-level patterns and risk factors associated with harmful behaviors. This capability represents a strategic shift from reactive responses to proactive prevention. By processing large-scale datasets derived from past incidents, such as the abuse reports studied by Puentes et al., these models can support authorities in identifying offender tactics and the underlying dynamics of criminal activity. These data-driven insights are essential for developing more targeted and effective preventive strategies, including awareness campaigns tailored to specific threats. This contributes to safer digital environments for children [6].

In a related study, Liu et al. [26] proposed an LSTM-based model that was trained using chat logs and enriched with sentiment vectors. The system could anticipate predatory behavior by analyzing the emotional tone of conversations. This provided a predictive layer that was highly relevant for early intervention efforts.

5.4. Comparative Efficiency of AI Methods Versus Traditional Techniques

Generative AI offers significant efficiency gains over traditional crime detection methods, particularly in handling large datasets and identifying complex behavioral patterns. Human capacity often limits traditional approaches, such as manual investigations and reactive victim reporting, leading to delayed responses and missed opportunities for early intervention. The efficiency gains of generative AI over traditional methods are further substantiated by comparative analyses focusing on threat detection and mitigation, which highlight how generative models can offer more dynamic and predictive capabilities compared to conventional AI systems [27].

In contrast, AI-driven systems can analyze massive amounts of data in real-time, scanning millions of online interactions simultaneously. For example, deep learning models such as BERT and Llama 2 achieve high accuracy rates (often exceeding 90%) in detecting grooming language, significantly outperforming traditional keyword-based filtering systems that typically struggle with subtle linguistic cues and coded language.

In addition, AI models trained on diverse datasets can detect grooming behavior at earlier stages, enabling proactive intervention before exploitation occurs. Studies comparing traditional methods with AI-based detection show a 40–60% improvement in early detection rates when using fine-tuned generative models.

However, these efficiency gains come with trade-offs. While AI systems excel at processing scale and speed, they are also prone to algorithmic bias and false positives, which can lead to the mistaken tagging of innocent users. This underscores the need for hybrid systems that combine AI automation with human oversight to ensure balanced and ethical decision-making.

In summary, generative AI holds transformative potential for combating pedophilia-related crimes in the digital sphere. Streamlining the detection of harmful material, uncovering predatory behavior, and predicting potential threats offers an unprecedented arsenal for child protection and crime prevention efforts.

6. Limitations and Ethical Risks of Using GenAI

Despite their potential, generative AI models present substantial risks when applied to sensitive areas like child abuse detection. This section critically examines the technical, ethical, and operational challenges identified in the reviewed studies. These challenges include privacy issues, bias, false positives, hallucinations, and the risk of misuse. The section highlights the importance of responsible deployment and robust safeguards.

There are many obstacles to harnessing the power of generative AI to prevent pedophilic crimes. Our review of the existing literature has identified the following limitations and potential risks: Privacy concerns, Algorithmic bias, Erroneous identifications, Misuse by malicious parties, and Hallucination. Beyond these specific technical and operational risks, broader societal concerns arise from the increasing interaction of children with generative AI tools, including potential psychological impacts and the facilitation of malicious uses such as sophisticated cyberbullying or grooming tactics, as highlighted in recent parliamentary reports [36].

6.1. Privacy Concerns

The use of AI systems to monitor digital interactions and online behavior raises significant concerns about the invasion of individual privacy. Striking a balance between protecting children and preserving individual privacy requires careful consideration and the establishment of strong safeguards. Without such safeguards, the ethical application of this technology risks being compromised and open to abuse [24]. One avenue being explored to address these privacy issues is the development of client-side detection mechanisms, where content analysis occurs directly on the user's device. Such approaches aim to provide strong privacy guarantees by minimizing data transmission to central servers, while still enabling the identification of harmful content at scale [37].

6.2. Algorithmic Bias

AI algorithms often absorb and replicate the biases embedded in their underlying training datasets, which can lead to biased results, including unfair profiling or misidentification of certain demographic groups. Addressing algorithmic bias is crucial to ensure the equitable operation of these systems and prevent the inappropriate targeting of individuals based on race, ethnicity, socioeconomic status, or other factors [30]. Addressing such algorithmic bias necessitates ongoing research into new methodologies for its detection and mitigation, alongside efforts to enhance algorithmic transparency, ensuring that AI systems operate equitably and their decision-making processes can be scrutinized [38]. The challenge of algorithmic bias is well-documented, with numerous research approaches focusing on its detection and mitigation through fairness-aware machine learning, adversarial debiasing, and post-hoc auditing, especially in critical domains such as criminal justice where such biases can have severe consequences [38].

6.3. Erroneous Identifications

The risk of erroneous conclusions, both false positives (incorrectly labeling innocent individuals as perpetrators) and false negatives (failing to identify true perpetrators), remains a persistent challenge. Mitigating these problems requires continuous refinement of AI models, exhaustive validation protocols, and vigilant human oversight to maintain accuracy, avoid false accusations, and maximize the detection of true threats [27].

6.4. Misuse by Malicious Parties

The prospect of AI tools being co-opted by malicious entities for nefarious purposes, such as creating deepfakes or manipulating evidence, is a serious risk. The threat of misuse is particularly acute with the rise of deepfake technology, where AI can be used to create highly realistic synthetic media for child exploitation, making detection and prevention significantly more challenging [39]. Recognizing this, governmental and academic collaborations are emerging to develop practical solutions and reusable datasets for evaluating deepfake detection tools, which is crucial given the scale of manipulated media investigators might face [40]. To prevent exploitation, robust regulatory frameworks and protective mechanisms are essential to ensure that the deployment and advancement of AI technologies are done responsibly and safely [6].

6.5. Hallucination

Like other artificial intelligence mechanisms, generative AI systems occasionally produce fabricated or illogical data, a phenomenon known as "hallucination". When applied to the detection of pedophilic activity, such inaccuracies could lead to unsubstantiated allegations or the failure to identify actual perpetrators. Continuous improvement of AI

models, coupled with vigilant human oversight, is essential to address this vulnerability and achieve reliable results [11].

To quantify these risks more concretely, recent studies have established formal benchmarks. A notable example is the HalluLens benchmark by Bang et al. [41], which provides a direct comparative evaluation of hallucination rates for leading LLMs. Their results reveal a wide variance in performance on its PreciseWikiQA task, with "hallucinated when not refused" rates ranging from as low as 26.84% for a Llama-3.1 model to as high as 85.22% for a Qwen2.5 model. This study also provides specific rates for leading commercial models, including GPT-4o (45.15%), Claude-3-Sonnet (56.24%), Mistral-7B (81.19%), and models from the Gemma family (68–76%) [41].

Mitigating these limitations and dangers requires relentless scrutiny, principled frameworks, and fortified safeguards to ensure the conscientious and effective use of generative AI in the crusade against pedophilic crimes. By skillfully navigating these nuances, we can unlock AI's immense potential to protect children, all while steadfastly preserving privacy, justice, and ethical accountability.

6.6. Legal and Workflow Integration Challenges

For evidentiary admissibility, the "black-box" nature of AI is a primary obstacle [42]. Since it is often impossible to retrace an algorithm's decision-making process, using its output as direct evidence in court is problematic [42]. The EUCPN paper recommends that AI results should be treated as "conjecture in the realm of probability", not as conclusive facts [42]. Therefore, workflows must ensure that humans are always the ultimate decision-makers regarding any intervention [42]. Furthermore, the proposed EU AI Act classifies predictive policing systems as "high-risk", subjecting them to strict transparency and human oversight requirements, which are prerequisites for legal acceptance [42].

Regarding GDPR compliance, any implementation must adhere to the EU's legal framework, including the GDPR and the Law Enforcement Directive (LED) [42]. The LED mandates that authorities conduct a Data Protection Impact Assessment (DPIA) before deployment to identify and mitigate risks to citizens' fundamental rights [42]. This is a critical step in any legally compliant workflow.

Jurisdictional challenges arise from the fragmented landscape of predictive policing within the EU. Different Member States, such as the Netherlands, Germany, and Estonia, use disparate systems based on unique national data sources and methodologies [42]. For instance, the Dutch Crime Anticipation System (CAS) utilizes demographic data, whereas Estonia's model incorporates border crossing information for person-based predictions [42]. This lack of a standardized approach creates significant challenges for developing harmonized, cross-border operational workflows, as an AI-generated risk score in one country may not be legally or methodologically compatible with those in another.

7. Synthesis of Findings

This section summarizes the key insights from the literature review and explains how generative AI models can help detect and prevent pedophilia-related crimes. Three overarching findings are highlighted: the potential of generative models, their effectiveness in detection and prevention, and the ethical concerns that must be addressed.

7.1. Potential of Generative AI

The research underscores the remarkable promise of generative AI, particularly on LLMs, in the fight against pedophile crime. LLMs have the ability to sift through massive amounts of textual and coded data, including chat logs, social media activity, and

forum discussions, to uncover patterns and anomalies that indicate grooming tactics, the distribution of CSAM, and other related criminal behavior [11].

7.2. Effective Detection and Prevention

The results demonstrate how technologies from the generative AI domain can streamline the identification of CSAM. This distinction is critical: while discriminative AI is used for the final classification task, it increasingly leverages the robust Transformer-based architectures pioneered by generative models, thereby enhancing its detection capabilities [10]. By examining communication patterns, linguistic cues, and digital traces, AI can help identify individuals engaged in grooming practices or attempting to exploit minors. Furthermore, AI models have the potential to anticipate likely offenders by analyzing their digital behavior and specific behavioral indicators [7].

7.3. Ethical Challenges and Mitigation Strategies

The literature highlights the ethical dilemmas associated with privacy concerns, algorithmic bias, and the risk of AI misuse in this sensitive context. To mitigate these issues, researchers advocate for measures such as thorough bias testing, promoting transparency and clarity in the operation of AI models, and creating well-defined legal and ethical frameworks for the development and deployment of these technologies [25].

8. Gaps and Future Research on GenAI for Child Safety

This section outlines priority areas for future research based on the gaps and challenges identified in the literature. These areas include developing adaptive and explainable models, creating robust datasets, devising cross-platform detection strategies, and integrating AI into human-centered intervention workflows.

While the promise of generative AI to combat the scourge of pedophilic crime is evident in the ever-evolving field of research, there are still numerous unexplored areas that warrant further investigation and development. In the following subsections, we propose five main topics.

8.1. Development of Dynamic and Adaptive Models

Research should focus on models that can continuously learn and adapt to new grooming tactics, evolving language (slang, emojis, coded communication), and changing platform characteristics.

8.2. Cross-Platform Threat Intelligence

Develop frameworks and AI models that can simultaneously ingest, fuse, and analyze data from multiple online platforms (social media, gaming chats, forums) to build comprehensive user profiles and detect coordinated or dispersed predatory activity.

8.3. Enhance Contextual Understanding with Advanced NLP

Explore more sophisticated NLP techniques, including graph neural networks for conversation structure, advanced attention mechanisms, and multimodal analysis (integrating text, image, video, and user metadata) to improve the understanding of conversational context and intent. The advancement of multimodal AI, while beneficial, also introduces new risks, such as the potential for harmful textual instructions to be embedded within image files, thereby evading traditional safety filters and posing new challenges for child protection that future research must address [43].

8.4. Robust, Realistic, and Bias-Mitigated Datasets

Collaborative efforts are needed to create larger, more diverse, and more realistic datasets that capture modern grooming behaviors across different platforms and demographics. Synthetic data generation techniques (with careful use of Generative AI) and robust bias detection/mitigation strategies during data collection and model training are critical. Given the inherent sensitivity and restricted access to actual CSAM, developing robust models is exceptionally challenging. Future work could benefit from establishing protocols that rely on proxy tasks, allowing for the design and preliminary validation of detection models without direct interaction with illicit material during the initial stages, thereby mitigating risks and ethical concerns associated with data handling [44].

8.5. Hybrid Human-AI Systems for Intervention

Investigate optimal ways to integrate AI detection tools into operational workflows. This includes designing interfaces for human analysts, establishing protocols for verifying AI flags, minimizing the impact of false positives, and studying the effectiveness of AI-assisted interventions versus manual methods.

8.6. Advancing Explainable AI for Generative Models

Advancing the explainability of generative AI models is a critical future research avenue. As current models often operate as 'black boxes', understanding their decision-making processes is paramount when deploying such systems for sensitive tasks like pedophilia detection. As highlighted by recent surveys, developing explainable AI techniques tailored for LLMs can help in fostering trust, ensuring accountability, and facilitating the identification of potential biases or failure modes [45].

9. Conclusions

Generative artificial intelligence, particularly LLMs, is emerging as a formidable ally in the fight against child exploitation. Their ability to process large amounts of information, recognize subtle patterns, and generate text that mimics human speech presents an extraordinary opportunity to automate the identification of CSAM, unmask online predators, and predict potential offenders. However, the use of generative AI in such a sensitive area requires careful attention to ethical dilemmas and potential dangers.

In order to guarantee responsible and effective implementation, it is imperative to protect privacy, mitigate algorithmic bias, and prevent exploitation by nefarious actors. The deployment of generative AI for child protection must also navigate and align with evolving legal frameworks. A landmark development in this area is the EU AI Act, formally adopted in 2024, which establishes a risk-based regulatory approach. Under this legislation, AI systems intended for use in law enforcement are classified as 'high-risk', mandating that they adhere to strict requirements regarding data quality, transparency, and fundamental rights. This reinforces the necessity for robust human oversight and the ethical safeguards discussed in this review, ensuring that these powerful tools are developed and deployed responsibly [46]. To ensure their practical applicability, these AI models should be developed in collaboration with child protection organizations, cybersecurity experts, and law enforcement.

Platforms that host user-generated content, such as social media networks and gaming communities, should integrate AI-driven moderation tools to detect grooming behavior while proactively adhering to strict privacy policies.

Resolving ethical dilemmas, developing real-time detection and intervention frameworks, and enhancing the capabilities of generative AI systems should be top priorities for future research. We can significantly improve child protection protocols and make

progress in the fight against pedophilic crimes in the digital age through the responsible and thoughtful use of generative AI.

Future work will incorporate valuable regional studies and gray literature from non-English sources, such as Portuguese, to broaden the scope and enhance the generalizability of the findings globally. While some studies mention aspects of computational cost and efficiency, as noted in Table 5 and Section 5.4, mention aspects of computational cost and efficiency, a holistic and standardized analysis of cost-effectiveness is not consistently available across the literature. Many studies focus on accuracy metrics but omit crucial details on computational complexity, hardware demands, inference latency, and scalability costs. Therefore, a systematic evaluation of these operational factors remains a critical gap. Future research should benchmark these AI models for both their detection accuracy and their practical and financial viability for law enforcement and child protection agencies.

Author Contributions: Conceptualization, F.S., J.B. and R.R.S.; methodology, F.S., J.B. and R.R.S.; investigation, F.S., J.B. and R.R.S.; writing—original draft preparation, F.S.; writing—review and editing, J.B. and R.R.S.; supervision, J.B. and R.R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BoW	Bag-of-Words
CSA	Child Sexual Abuse
CSAM	Child Sexual Abuse Material
FL	Federated Learning
GenAI	General Artificial Intelligence
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
LSTM-RNN	Long Short-Term Memory Recurrent Neural Networks
ML	Machine Learning
NLP	Natural Language Processing
OCR	Optical Character Recognition
PCI	Predatory Conversation Identification
SGS-V	Sexual Grooming Scale Victim Version
SimCSE	Simple Contrastive Sentence Embedding
SPI	Sexual Predator Identification
SVM	Support Vector Machine
VPD	Victim-Predator Differentiation

References

1. Marvasti, J.A. (Ed.) *Psychiatric Treatment of Sexual Offenders: Treating the Past Traumas in Traumatizers. A Bio-Psycho-Social Perspective*; Charles C Thomas Publisher: Springfield, IL, USA, 2004.
2. Cook, D.; Zilka, M.; DeSandre, H.; Giles, S.; Weller, A.; Maskell, S. Can We Automate the Analysis of Online Child Sexual Exploitation Discourse? *arXiv* **2022**, arXiv:2209.12320.
3. Wolbers, H.; Cubitt, T.; Cahill, M.J. Artificial intelligence and child sexual abuse: A rapid evidence assessment. *Trends Issues Crime Crim. Justice* **2025**, *711*, 1–18. [CrossRef]
4. Levy, I.; Robinson, C. Thoughts on child safety on commodity platforms. *arXiv* **2022**, arXiv:2207.09506.

5. UNICRI—United Nations Interregional Crime and Justice Research Institute. New! How AI Is Leading the Fight Against Online Child Abuse. 2023. Available online: https://unicri.org/News/AI-for-Safer-Children-%20article-Emerging-Europe (accessed on 6 May 2025).
6. Puentes, J.; Castillo, A.; Osejo, W.; Calderón, Y.; Quintero, V.; Saldarriaga, L. Guarding the Guardians: Automated Analysis of Online Child Sexual Abuse. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2–6 October 2023; pp. 3730–3734. [CrossRef]
7. Borj, P.R.; Raja, K.; Bours, P. Online grooming detection: A comprehensive survey of child exploitation in chat logs. *Knowl.-Based Syst.* **2023**, *259*, 110039. [CrossRef]
8. Mou, J.; Duan, P.; Gao, L.; Liu, X.; Li, J. An effective hybrid collaborative algorithm for energy-efficient distributed permutation flow-shop inverse scheduling. *Future Gener. Comput. Syst.* **2022**, *128*, 521–537. [CrossRef]
9. Jeglic, E.L.; Winters, G.M.; Johnson, B.N. Identification of red flag child sexual grooming behaviors. *Child Abus. Negl.* **2023**, *136*, 105998. [CrossRef] [PubMed]
10. Rani, G.; Singh, J.; Khanna, A. Comparative Analysis of Generative AI Models. In Proceedings of the 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT), Faridabad, India, 23–24 November 2023; pp. 760–765. [CrossRef]
11. Prosser, E.; Edwards, M. Helpful or Harmful? Exploring the Efficacy of Large Language Models for Online Grooming Prevention. In Proceedings of the EICC 2024: European Interdisciplinary Cybersecurity Conference, Xanthi, Greece, 5–6 June 2024; pp. 1–10. [CrossRef]
12. Rane, N.; Choudhary, S.; Rane, J. Gemini versus ChatGPT: Applications, performance, architecture, capabilities, and implementation. *J. Appl. Artif. Intell.* **2024**, *5*, 69–93. [CrossRef]
13. Buscemi, A.; Proverbio, D. ChatGPT vs Gemini vs LLaMA on Multilingual Sentiment Analysis. *arXiv* **2024**, arXiv:2402.01715.
14. Pasca, P.; Signore, F.; Tralci, C.; Longo, M.; Preite, G.; Ciavolino, E. Detecting online grooming at its earliest stages: Development and validation of the Online Grooming Risk Scale. *Mediterr. J. Clin. Psychol.* **2022**, *10*, 1–24. [CrossRef]
15. McGhee, I.; Bayzick, J.; Kontostathis, A.; Edwards, L.; McBride, A.; Jakubowski, E. Learning to Identify Internet Sexual Predation. *Int. J. Electron. Commer.* **2011**, *15*, 103–122. [CrossRef]
16. Upadhyay, A.; Chaudhari, A.; Arunesh; Ghale, S.; Pawar, S.S. Detection and prevention measures for cyberbullying and online grooming. In Proceedings of the 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 19–20 January 2017. [CrossRef]
17. Keyvanpour, M.; Nayebi, N.G.; Ebrahimi, M.; Ormandjieva, O.; Suen, C.Y. Automated identification of child abuse in chat rooms by using data mining. In *Data Mining Trends and Applications in Criminal Science and Investigations*; IGI Global: Hershey, PA, USA, 2016; pp. 245–274. [CrossRef]
18. Ngejane, C.H.; Mabuza-Hocquet, G.; Eloff, J.H.P.; Lefophane, S. Mitigating Online Sexual Grooming Cyber-crime on Social Media Using Machine Learning: A Desktop Survey. In Proceedings of the 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 6–7 August 2018. [CrossRef]
19. Anderson, P.; Zuo, Z.; Yang, L.; Qu, Y. An Intelligent Online Grooming Detection System Using AI Technologies. In Proceedings of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 23–26 June 2019. [CrossRef]
20. Pranoto, H.; Gunawan, F.E.; Soewito, B. Logistic Models for Classifying Online Grooming Conversation. *Procedia Comput. Sci.* **2015**, *59*, 357–365. [CrossRef]
21. Fauzi, M.A.; Wolthusen, S.; Yang, B.; Bours, P.; Yeng, P. Identifying Sexual Predators in Chats Using SVM and Feature Ensemble. In Proceedings of the 2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Windhoek, Namibia, 16–18 August 2023; pp. 70–75. [CrossRef]
22. Ebrahimi, M.; Suen, C.Y.; Ormandjieva, O. Detecting predatory conversations in social media by deep Convolutional Neural Networks. *Digit. Investig.* **2016**, *18*, 33–49. [CrossRef]
23. Nasir, L.H.M.; Saaya, Z.; Baharon, M.R. Identifying Online Sexual Grooming Content in Social Media Using Classification Technique. *J. Adv. Comput. Technol. Appl.* **2022**, *4*, 33–42.
24. Mila, K.C.; Montreal, H.; Caporossi, G.; Rabbany, R.; De Cock, M.; Mila, G.F. Early Detection of Sexual Predators with Federated Learning. May 2023. Available online: https://openreview.net/pdf?id=M84OnT0ZvDq (accessed on 6 May 2025).
25. Eilifsen, T.N.; Shrestha, B.; Bours, P. Early Detection of Cyber Grooming in Online Conversations: A Dynamic Trust Model and Sliding Window Approach. In Proceedings of the 2023 21st International Conference on Emerging eLearning Technologies and Applications (ICETA), Stary Smokovec, Slovakia, 26–27 October 2023; pp. 129–134. [CrossRef]
26. Liu, D.; Suen, C.Y.; Ormandjieva, O. A Novel Way of Identifying Cyber Predators. *arXiv* **2017**, arXiv:1712.03903.
27. Vogt, M.; Leser, U.; Akbik, A. Early detection of sexual predators in chats. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 1–6 August 2021; pp. 4985–4999. [CrossRef]

28. Simen, M.A. Fine Tuning BERT for Detecting Cyber Grooming in Online Chats. Master's Thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2023. Available online: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3088473 (accessed on 6 May 2025).
29. Borj, P.R.; Raja, K.; Bours, P. Detecting Online Grooming by Simple Contrastive Chat Embeddings. In Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy, Charlotte, NC, USA, 26 April 2023; pp. 57–65. [CrossRef]
30. Nguyen, T.T.; Wilson, C.; Dalins, J. Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 9–11 October 2024; pp. 613–618. [CrossRef]
31. Kim, S.; Lee, B.; Maqsood, M.; Moon, J.; Rho, S. Deep Learning-Based Natural Language Processing Model and Optical Character Recognition for Detection of Online Grooming on Social Networking Services. *Comput. Model. Eng. Sci.* **2025**, *143*, 2079–2108. [CrossRef]
32. Kim, D.; Kim, T.; Yang, J. Early Detection of Online Grooming with Language Models. In Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, Catania International Airport, Catania, Italy, 31 March–4 April 2025; pp. 963–970. [CrossRef]
33. Hamm, L. Advancing Grooming Detection in Chat Logs: Comparing Traditional Machine Learning and Large Language Models with a Focus on Predator Tone. Master's Thesis, Uppsala University, Uppsala, Sweden, 2025. Available online: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-550352 (accessed on 6 May 2025).
34. Pereira, M.; Dodhia, R.; Anderson, H.; Brown, R. Metadata-Based Detection of Child Sexual Abuse Material. *IEEE Trans. Dependable Secur. Comput.* **2023**, *21*, 3153–3164. [CrossRef]
35. Struckman, K. Wilson Center. Combatting AI-Generated CSAM. February 2023. Available online: https://www.wilsoncenter.org/article/combatting-ai-generated-csam (accessed on 6 May 2025).
36. European Parliament. Children and Generative AI. February 2025. Available online: https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2025)769494 (accessed on 6 May 2025).
37. Hua, Y.; Namavari, A.; Cheng, K.; Naaman, M.; Ristenpart, T. Increasing Adversarial Uncertainty to Scale Private Similarity Testing. In Proceedings of the 31st USENIX Security Symposium, Boston, MA, USA, 10–12 August 2022; pp. 1777–1794. Available online: https://www.usenix.org/system/files/sec22summer_hua.pdf (accessed on 6 May 2025).
38. Liang, P.P.; Wu, C.; Morency, L.P.; Salakhutdinov, R. Towards Understanding and Mitigating Social Biases in Language Models. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; Volume 139, pp. 6565–6576. Available online: https://arxiv.org/pdf/2106.13219 (accessed on 6 May 2025).
39. Monash University. Digital Child Abuse: Deepfakes and the Rising Danger of AI-Generated Exploitation. February 2025. Available online: https://lens.monash.edu/@politics-society/2025/02/25/1387341/digital-child-abuse-deepfakes-and-the-rising-danger-of-ai-generated-exploitation (accessed on 6 May 2025).
40. GOV.UK. Innovating to Detect Deepfakes and Protect the Public. February 2023. Available online: https://www.gov.uk/government/case-studies/innovating-to-detect-deepfakes-and-protect-the-public (accessed on 6 May 2025).
41. Bang, Y.; Ji, Z.; Schelten, A.; Hartshorn, A.; Fowler, T.; Zhang, C.; Cancedda, N.; Fung, P. HalluLens: LLM Hallucination Benchmark. *arXiv* **2025**, arXiv:2504.17550. Available online: https://arxiv.org/pdf/2504.17550 (accessed on 6 May 2025).
42. EUCPN—European Crime Prevention Network. Artificial Intelligence and Predictive Policing: Risks and Challenges. June 2022. Available online: https://www.eucpn.org/document/recommendation-paper-artificial-intelligence-and-predictive-policing-risks-and-challenges (accessed on 6 May 2025).
43. GlobeNewswire. Multimodal AI at a Crossroads: Report Reveals CSEM Risks. May 2025. Available online: https://www.globenewswire.com/news-release/2025/05/08/3077301/0/en/Multimodal-AI-at-a-Crossroads-Report-Reveals-CSEM-Risks.html (accessed on 6 May 2025).
44. Coelho, T.; Ribeiro, L.S.F.; Macedo, J.; Santos, J.A.D.; Avila, S. Minimizing Risk Through Minimizing Model-Data Interaction: A Protocol For Relying on Proxy Tasks When Designing Child Sexual Abuse Imagery Detection Models. *arXiv* **2025**, arXiv:2505.06621. Available online: https://arxiv.org/pdf/2505.06621v1 (accessed on 6 May 2025).
45. Bilal, A.; Ebert, D.; Lin, B. LLMs for Explainable AI: A Comprehensive Survey. *arXiv* **2025**, arXiv:2504.00125.
46. European Parliament. Artificial Intelligence Act: MEPs Adopt Landmark Law. Press Release, 13 March 2024. Available online: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law (accessed on 6 May 2025).