Viewpoint

Leveraging AI to Investigate Child Maltreatment Text Narratives: Promising Benefits and Addressable Risks

Wilson Lukmanjaya[1], BSc; Tony Butler[1], BSc, MSc, PhD; Sarah Cox[2], BSc, PhD; Oscar Perez-Concha[3], BSc, PhD; Leah Bromfield[2], BSc, PhD; George Karystianis[1], BSc, MSc, PhD

[1]School of Population Health, University of New South Wales Sydney, Sydney, Australia
[2]Australian Center for Child Protection, University of South Australia, Adelaide, Australia
[3]Centre for Big Data Resarch in Health, University of New South Wales Sydney, Sydney, Australia

Corresponding Author:

Wilson Lukmanjaya, BSc
School of Population Health
University of New South Wales Sydney
F25, Samuels Building, Samuel Terry Ave, Kensington
Sydney 2033
Australia
Phone: 61 293853136
Fax: 61 93136185
Email: w.lukmanjaya@unsw.edu.au

Abstract

The trove of information contained in child maltreatment narratives represents an opportunity to strengthen the evidence base for policy reform in this area, yet it remains underutilized by researchers and policy makers. Current research into child maltreatment often involves the use of qualitative methodologies or structured survey data that are either too broad or not representative, thereby limiting the development of effective policy responses and intervention strategies. Artificial intelligence (AI) approaches such as large language models (AI models that understand and generate language) can analyze large volumes of child maltreatment narratives by extracting population-level insights on factors of interest such as mental health and treatment needs. However, when applying such methods, it is useful to have a framework on which to base approaches to the data. We propose a seven step framework: (1) data governance; (2) researcher vetting; (3) data deidentification; (4) data access; (5) feasibility testing of baseline methods; (6) large-scale implementation of black box algorithms; and (7) domain expert result validation for such exercises to ensure careful execution and limit the risk of privacy and security breaches, bias, and unreliable conclusions.

JMIR Pediatr Parent 2025;8:e73579; doi: 10.2196/73579

Keywords: artificial intelligence; large language models; child maltreatment narratives; text mining; benefits and risks; research framework

Smarter Tech, Safer Kids

The World Health Organization estimates around 1 in 2 children worldwide experience physical, sexual, and emotional abuse and neglect [1]. In Australia, the monitoring of child maltreatment (ie, "all forms of physical and emotional ill-treatment, sexual abuse, neglect, and exploitation that results in actual or potential harm to the child's health, development or dignity" [2,3]) relies on administrative data provided by child protection departments for local and national reporting purposes [4-6]. However, the structured nature of these data limits the reporting of more nuanced information such as maltreatment types, treatment needs, and mental health, preventing the ability to answer key questions. For example, the 2021 Australian Child Maltreatment Study [6] recorded the prevalence of physical abuse and sexual abuse at 32.0% and 28.5%, respectively, whereas the Australian Bureau of Statistics [5] in the same year reported that 1 in 7 (14%) children experienced physical and sexual abuse. Several factors contribute to such discrepancies, including the use of different methodologies (eg, self-report, cross-sectional surveys, and surveillance through administrative data), varying definitions of abuse and age ranges, and biases due to the small sample sizes and response rates [7-9]. While high-level reporting is important, both sources fail to provide a comprehensive picture of child maltreatment

necessary for effective policy responses. Characteristics such as uncommon maltreatment subtypes and treatment needs are lacking. Indeed, the Australian Productivity Commission states that "with technological developments and advances in analytical techniques, not only is the volume of data being generated and collected growing, but so too is the scope to make use of data in innovative ways in every sphere of life" [10]. Child protection should be no exception to this.

Research in child maltreatment has predominantly utilized statistical and artificial intelligence (AI) approaches such as artificial neural networks and predictive risk modeling applied to structured child protection administrative data [11-15]. Studies on health administrative records focused on developing models to predict cases of sexual abuse [11], physical abuse [12], and overall maltreatment [13-15], with robust performance. Furthermore, information has been successfully extracted using text mining techniques such as word embeddings, bag-of-words, and feature selection from free-text health notes [16-19] to detect physical abuse.

Nevertheless, rarely tapped sources of unstructured information in this area are child maltreatment narratives containing detailed descriptive accounts of incidents or experiences of physical or emotional harm to a child. These document the initial reasons for coming to the attention of authorities (ie, intake report), descriptions of a child's experience of abuse or neglect, psychologists' assessments, police and medical records, and a detailed history of child protection involvement. They also describe abuse types, substance use and abuse, recorded injuries, mental and behavioral problems, physical health conditions, parental and family characteristics, and the types and conditions of the accommodation. Utilizing the information contained within the narratives could fill existing knowledge gaps by highlighting less common abusive behaviors (eg, social abuse, coercive control) and enriching definitions and measures of abuse types and mental illnesses [20-22]. A limited number of studies have used these narratives to automatically extract information. Victor et al [23] and Perron et al [24,25] applied machine learning algorithms such as random forest and k-nearest neighbors on caseworker summaries to identify domestic violence events and opioid-related maltreatment risks, respectively. In addition, Saxena et al [26] used topic modeling to explore the concept and definitions of "risk" on child maltreatment records with promising performance. Despite the advent of newer technologies to conduct text analysis such as large language models (LLMs; AI models that can understand and generate language), limited attempts have been made [27,28] to implement this approach. In particular, Perron et al [27] reported an F_1-score of 89%-95% for their model to classify and extract substance-related problems, while more recently, Stoll et al [28] noted their approach to classify maltreatment subtype yielded a superior performance against human reviewers by 10%.

However, automatically mining information from child maltreatment records faces significant limitations [23-28]. Most importantly, the lack of sufficient training data restricts the AI model's capacity to learn complex patterns [29] such as nuanced indicators of abuse. In addition, the limited contextual understanding (even when using LLMs), along with the inconsistencies in narrative style and language, as caseworkers may document—or omit—events of child maltreatment through various terms, descriptions, and writing style, can lead to inaccurate outputs and falsehoods [26]. This is particularly critical when analyzing data related to Indigenous populations requiring a specific cultural lens through which to examine the data. Relying exclusively on one data type (eg, administrative data) can also skew the results and present an inaccurate picture [30].

Benefits Versus Risks

To manually process such voluminous and diverse information is extremely time consuming and not without significant challenges, including a lack of human resources to inspect large scale data and the associated high cost. AI pipelines, particularly applications of LLMs, offer a solution for efficiently processing vast amounts of child maltreatment narratives, comprehending the text for a broader range of tasks (ie, summarization, classification) [27,28,31], and offering potentially greater quality control due to its standardized approach compared to other (eg, qualitative) methods [32]. A primary concern of child protection services is to safeguard and monitor those in their care who may be at risk of maltreatment and neglect. Timely processing of many thousands of cases provides the possibility for professionals and policy makers to make faster decisions (eg, identifying clusters of abuse or mental illness using geographic areas) by predicting the severity of future abuse or allocating resources to high-prevalence communities and intervene earlier. Consequently, new evidence-based protocols for early intervention and prevention can be developed aimed at improving outcomes for children and families.

While appealing, AI applications come with security risks (eg, malware, data leakage), performance issues (eg, "AI hallucinations"), technical limitations (eg, black box bias), and ethical conundrums (eg, machine vs human decisions). Hidden malware or malicious code within unvetted AI algorithms with built-in third-party components could result in leaking sensitive information residing in child maltreatment narratives [33,34]. Child protection departments are extremely cautious about providing data access (making this data some of the most difficult to access) as children's data privacy risks could expose them to further harm. When the data involve Indigenous populations, additional cultural considerations for access, governance, and interpretation are required. Even with data access approval, challenges related to narrative data quality, such as data omission, will persist as researchers have limited control over these factors. Cultural and linguistic biases, as well as selection bias in the training data, also pose risks for misinterpretation of findings. In particular, LLM performance can be vulnerable due to "hallucinations" or misinformation with models generating inaccurate or false conclusions. This is further amplified by black box bias [35-37], since the decision-making process is not visible or understandable by humans and, therefore, it is hard to conduct any detailed error analysis.

For these reasons, AI-generated misinformation can have larger implications due to false positives and negatives, fabrication of summarization, and misclassifications. In child maltreatment cases, this could lead to the misidentification of mental health disorders, abuse and neglect misclassification, and unnecessary family separation [38-42]. Falsehoods could be interpreted as gold standards. Therefore, child maltreatment policies require a decision framework that draws conclusions from both machines and humans, with the final decision relying on human expertise. While some may propose automated approaches can reduce human errors, bias, and the time required to process reports, replacing human decision-making with AI-powered approaches requires ethical protocols to be in place. Relying exclusively on algorithms and their outputs should be discouraged as they lack the complexities and subtleties of human reasoning and decision making [43-47]. Human judgement needs to take priority over AI outputs [48] and adhere to research ethics and legal standards [49-51]. Machine-made choices bear potentially severe consequences when it comes to enhancing the evidence base and policy translation leading to social injustice and population oversurveillance [48]. Considering the inherent risk of biases [52,53] (ie, underrepresented groups sampling bias) and the cost of hardware requirements, security protocols, and manual code inspection, it is difficult to apply such algorithms widely without the appropriate precautions.

Risk mitigation strategies are essential to ensure the appropriate application of automated approaches in areas such as child protection. First, using deidentified data and downloadable models from verified sources minimizes the risk of leaking private information. The incorporation of local LLMs could be set up using a virtual machine and offline to prohibit data access and the possibility of reidentifying personal records if information were to be siphoned off through hidden malware. False outcomes could be mitigated through improved quality of input data, further training to improve the model's understanding and designing clear, precise instructions for the model (ie, design prompt) that steer its response towards accurate information [47,48]. Governance infrastructures are also necessary to ensure appropriate handling of such sensitive data and the generated findings, especially if these data involve Indigenous populations, which would require input from key stakeholder groups and people with lived experience to ensure culturally sensitive analysis and interpretation. Finally, pipelines need to have expert-in-the-loop architecture to improve the quality and reliability of results. Ensuring that pipelines powered up by LLMs operate transparently with domain experts' input not only protects against biased decision-making but also aligns with efforts to safeguard personal data, as some algorithms can obscure the processes that may compromise fairness and privacy.

Proposed Framework

Existing research using AI approaches on child maltreatment narratives follows the establishment of the technical pipeline to extract information and obtain data access. Additional steps (eg, governance, data deidentification, researcher vetting) are also required to not only enhance the performance of such approaches but to limit security risks, ensure translatable outputs, and boost public confidence in such approaches. For these reasons, we propose a seven-step framework for processing child maltreatment narratives with LLMs or any black box algorithm: (1) data governance (including cultural considerations where appropriate), (2) researcher vetting, (3) data deidentification, (4) data access, (5) feasibility testing of baseline methods, (6) large-scale implementation of black box algorithms, and (7) domain expert result validation. These steps align with Gillingham's [48] principles of algorithmic accountability in social work and existing policy guidelines on AI for children [54,55] that prioritize children's safety, well-being, data privacy, and AI transparency. Although Gillingham's principles are rooted in decision support tools, they are relevant due to the generated results potentially contributing to policy making and translation.

The seven steps include:

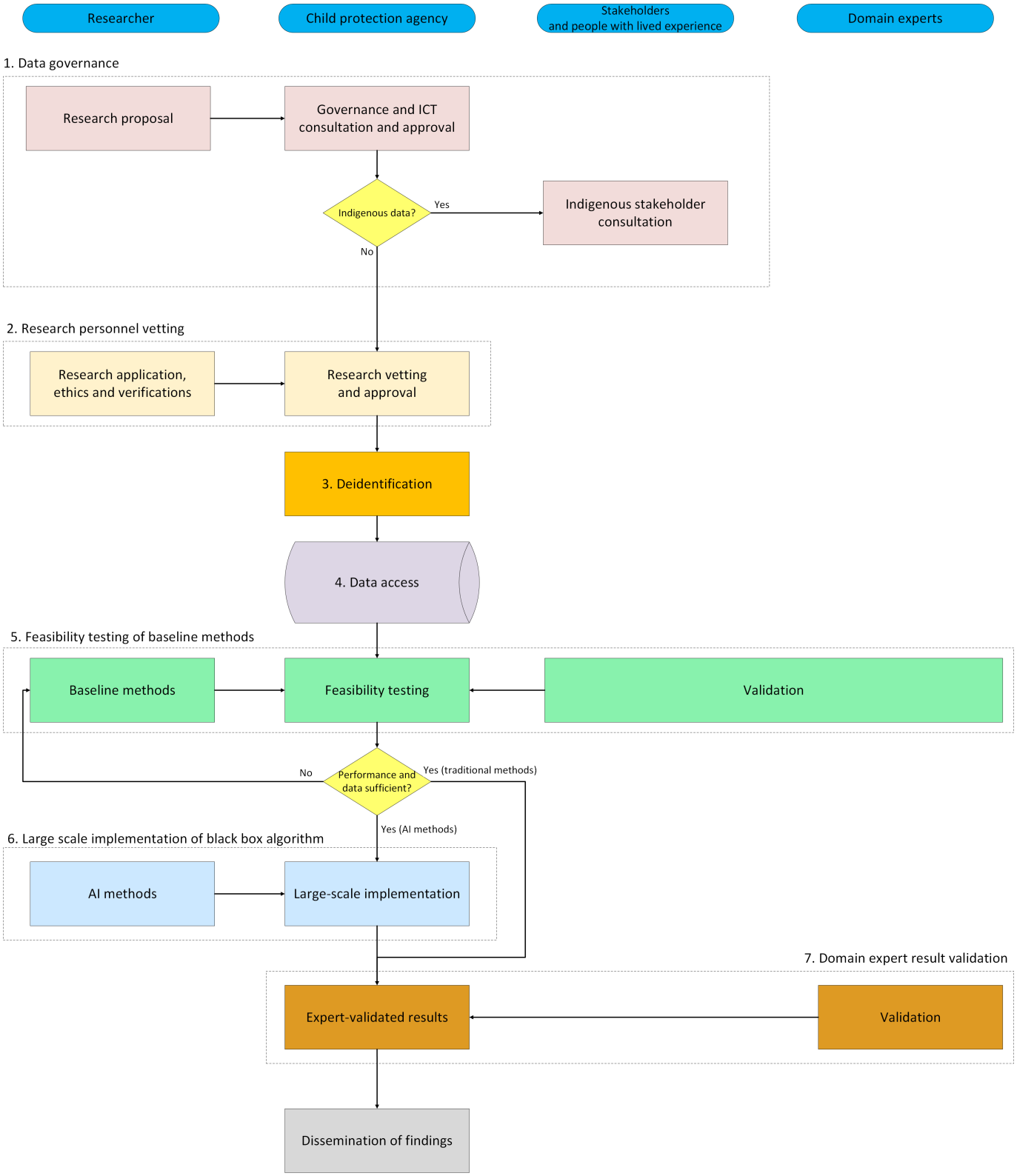
1. Data governance: this foundational step focuses on establishing appropriate data governance infrastructures on both the child maltreatment data and the generated outputs. Key stakeholder agencies such as departments of child protection and information and communication technology services along with ethics committees must be engaged before any data analysis takes place. If the data relate to Indigenous populations, it is necessary to include relevant community representatives to ensure the analysis interpretation and dissemination of findings are culturally appropriate and align with Indigenous data sovereignty.
2. Researcher vetting: researchers need to be vetted (eg, have no track record of criminal activity, have previous experience with handling and analyzing sensitive data, be domestically based) to ensure responsible research, including a working with children government check. Signing a nondisclosure agreement can further minimize potential risks such as data privacy breaches and identity theft.
3. Data deidentification: as child protection data contains extremely sensitive information (eg, name, addresses, date of birth), to minimize privacy breaches, deidentification of child maltreatment narratives is recommended. Thus, in the unlikely event of data leakage, deidentification ensures that no identifiable information is exposed in the dataset.
4. Data access: accessing child protection data should be done within the vicinity of child protection agencies who can assign researchers limited access and maintain history logs of accessed or modified documents. Such an approach will ensure transparency and provide the capacity for tracking errors and auditability.
5. Feasibility testing of baseline methods: conducting a feasibility test before any large-scale implementation is important. Implementing simpler approaches that rely on lexical pattern identification (eg, rule-based) or traditional machine learning algorithms offers a controlled, auditable environment with enhanced

security reducing the risk of unintended data exposure, allowing for more interpretability on potential errors (ie, false positives, false negatives) without requiring an extraneous number of resources and costs [56].

6. Large scale implementation of black box algorithms: the implementation of black box approaches or algorithms ensures efficient and effective large-scale processing of voluminous datasets. Evaluation metrics such as precision and recall ensure robust performance and support scalable research to address knowledge gaps in child maltreatment. Models such as LLMs can assist in the contextual understanding and scalability of the research to ultimately narrow the information gap within specific research aims in child maltreatment. These applications can be applied offline through a secure virtual machine environment. Through this approach, child protection organizations can ensure that researchers have no means of extracting and downloading identifiable data without explicit approval from participating governmental departments.
7. Domain expert result validation: the returned results need to be validated by domain experts in child maltreatment to ensure that the extracted or generated information does not lead to inaccurate or misleading conclusions. This can be done by manually checking a random sample of narratives and conducting interannotator agreement (via metrics such as Cohen κ coefficient) between 2 to 3 experts. Disagreements between domain experts regarding claims may require researchers to retrain and re-evaluate their method. This evaluation and validation process will ensure the findings' accuracy and relevancy are aligned with ethical standards such as the policy guidance for AI for children [54] or transnational frameworks for AI and children's rights and well-being [55], ensuring accuracy and fairness and limiting risks of potential harm before being used towards shaping existing or new policies and protocols.

Our proposed 7-step framework offers a novel research pipeline specific to ethical and secure use of AI in child maltreatment research. By introducing data governance, the inclusion of stakeholders with lived experience and structured deidentification processes and feasibility testing, the framework prioritizes both ethical and cultural safety while promoting analytical rigor. Incorporating domain expert validation further enhances the interpretability, contextual relevance, and ethical soundness of AI outputs—elements that are largely absent in the current literature. An illustrative example of a brief case scenario is presented in Figure 1. While it is understood that the framework may not apply to all contexts—given that protocols can vary across countries and may not have Indigenous cultures—the core principles are common. Adhering to these principles supports data privacy and cultural sensitivity and helps mitigate the uncaptured context problem and bias mitigation within the field through domain expert result validation [57].

Figure 1. Illustrative example of the proposed 7-step framework for responsible AI use in child maltreatment research. AI: artificial intelligence; ICT: information and communications technology.



Conclusions

As the complexity and scale of child maltreatment increases and the volume of information expands, an opportunity exists to apply new and sophisticated approaches that rely on AI to strengthen the evidence base by processing large volumes of narrative data. However, with this opportunity come risks that need to be minimized. The implementation of AI methods, particularly black box methods such as LLMs, should follow a framework of seven key steps—(1) data governance, (2) researcher vetting, (3) data deidentification, (4) data access, (5) feasibility testing of baseline methods, (6) large-scale implementation of black box algorithms, and (7)

domain expert result validation—to ensure culturally sensitive governance, minimization of privacy risks, improvement data analysis, and optimization of AI usability in strengthening the evidence base for child maltreatment. Further research is needed to evaluate the framework's effectiveness and practice to improve policies and lead to better outcomes in child protection.

Data Availability

No generative artificial intelligence was used in any portion of the manuscript generation.

Authors' Contributions

WL conceived the study and conducted the drafting, literature review, revision, and submission of the manuscript. OP-C, LB, and SC drafted and revised the study. TB helped with the study design, revision, and supervision. GK helped with study conception and design, revision, and supervision.

Conflicts of Interest

None declared.

References

1. Hillis S, Mercy J, Amobi A, Kress H. Global prevalence of past-year violence against children: a systematic review and minimum estimates. Pediatrics. Mar 2016;137(3):e20154079. [doi: 10.1542/peds.2015-4079] [Medline: 26810785]
2. Daley SF, Gonzalez D, Bethencourt Mirabal A, Afzal M. Child abuse and neglect. In: StatPearls. StatPearls Publishing; 2025. [Medline: 29083602]
3. Child maltreat. World Health Organization. 2024. URL: https://www.who.int/news-room/fact-sheets/detail/child-maltreatment [Accessed 2025-04-22]
4. Child protection Australia 2021-22. Australian Institute of Health and Welfare. 2023. URL: https://www.aihw.gov.au/reports/child-protection/child-protection-australia-2021-22 [Accessed 2024-04-16]
5. 1 in 7 Australians have experienced childhood abuse. Australian Bureau of Statistics. 2023. URL: https://www.abs.gov.au/media-centre/media-releases/1-7-australians-have-experienced-childhood-abuse [Accessed 2024-04-16]
6. Mathews B, Pacella R, Dunne M, et al. The Australian Child Maltreatment Study (ACMS): protocol for a national survey of the prevalence of child abuse and neglect, associated mental disorders and physical health problems, and burden of disease. BMJ Open. May 11, 2021;11(5):e047074. [doi: 10.1136/bmjopen-2020-047074] [Medline: 33980529]
7. Mathews B, Pacella R, Scott JG, et al. The prevalence of child maltreatment in Australia: findings from a national survey. Med J Aust. Apr 3, 2023;218 Suppl 6(Suppl 6):S13-S18. [doi: 10.5694/mja2.51873] [Medline: 37004184]
8. 4906055003 - personal safety survey, Australia: user guide, 2016: methodology. Australian Bureau of Statistics. 2017. URL: https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4906.0.55.003main+features222016 [Accessed 2025-04-22]
9. 4906.0.55.003 - personal safety survey, Australia: user guide, 2016: response rates. Australian Bureau of Statistics. 2017. URL: https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4906.0.55.003main+features242016 [Accessed 2025-04-22]
10. Data availability and use. Productivity Commission. 2017. URL: https://www.pc.gov.au/inquiries/completed/data-access/report/data-access.pdf [Accessed 2024-04-16]
11. Ucuz I, Ari A, Ozcan OO, Topaktas O, Sarraf M, Dogan O. Estimation of the development of depression and PTSD in children exposed to sexual abuse and development of decision support systems by using artificial intelligence. J Child Sex Abus. Jan 2022;31(1):73-85. [doi: 10.1080/10538712.2020.1841350] [Medline: 33206583]
12. Shahi N, Shahi AK, Phillips R, Shirek G, Lindberg DM, Moulton SL. Using deep learning and natural language processing models to detect child physical abuse. J Pediatr Surg. Dec 2021;56(12):2326-2332. [doi: 10.1016/j.jpedsurg.2021.03.007] [Medline: 33838900]
13. Vaithianathan R, Maloney T, Putnam-Hornstein E, Jiang N. Children in the public benefit system at risk of maltreatment: identification via predictive modeling. Am J Prev Med. Sep 2013;45(3):354-359. [doi: 10.1016/j.amepre.2013.04.022] [Medline: 23953364]
14. Horikawa H, Suguimoto SP, Musumari PM, Techasrivichien T, Ono-Kihara M, Kihara M. Development of a prediction model for child maltreatment recurrence in Japan: a historical cohort study using data from a Child Guidance Center. Child Abuse Negl. Sep 2016;59:55-65. [doi: 10.1016/j.chiabu.2016.07.008] [Medline: 27517122]
15. Choi J, Kim K. Predictive risk modeling for recurrence of child maltreatment using cases from the National Child Maltreatment Data System in Korea: exploratory data analysis using data mining algorithm. Prev Sci. Nov 2022;23(8):1517-1530. [doi: 10.1007/s11121-022-01446-5] [Medline: 36251208]
16. Amrit C, Paauw T, Aly R, Lavric M. Identifying child abuse through text mining and machine learning. Expert Syst Appl. Dec 2017;88:402-418. [doi: 10.1016/j.eswa.2017.06.035]
17. Jadhav P, Sears T, Floan G, et al. Application of a machine learning algorithm in prediction of abusive head trauma in children. J Pediatr Surg. Jan 2024;59(1):80-85. [doi: 10.1016/j.jpedsurg.2023.09.027] [Medline: 37858394]

18. Annapragada AV, Donaruma-Kwoh MM, Annapragada AV, Starosolski ZA. A natural language processing and deep learning approach to identify child abuse from pediatric electronic medical records. PLoS ONE. 2021;16(2):e0247404. [doi: 10.1371/journal.pone.0247404] [Medline: 33635890]
19. Tiyyagura G, Asnes AG, Leventhal JM, et al. Development and validation of a natural language processing tool to identify injuries in infants associated with abuse. Acad Pediatr. Aug 2022;22(6):981-988. [doi: 10.1016/j.acap.2021.11.004] [Medline: 34780997]
20. Laajasalo T, Cowley LE, Otterman G, et al. Current issues and challenges in the definition and operationalization of child maltreatment: a scoping review. Child Abuse Negl. Jun 2023;140:106187. [doi: 10.1016/j.chiabu.2023.106187] [Medline: 37030235]
21. Brewsaugh K, Holmes AK, Richardson A, et al. Research and knowledge gaps in child welfare in the United States: a national survey of agency staff, allied disciplines, tribal leaders, and people who have experienced child welfare. Child Youth Serv Rev. Jul 2022;138:106496. [doi: 10.1016/j.childyouth.2022.106496]
22. Higgins DJ, Mathews B, Pacella R, et al. The prevalence and nature of multi-type child maltreatment in Australia. Med J Aust. Apr 3, 2023;218 Suppl 6(Suppl 6):S19-S25. [doi: 10.5694/mja2.51868] [Medline: 37004183]
23. Victor BG, Perron BE, Sokol RL, Fedina L, Ryan JP. Automated identification of domestic violence in written child welfare records: leveraging text mining and machine learning to enhance social work research and evaluation. J Soc Social Work Res. Dec 1, 2021;12(4):631-655. [doi: 10.1086/712734]
24. Perron BE, Victor BG, Bushman G, et al. Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning. Child Abuse Negl. Dec 2019;98:104180. [doi: 10.1016/j.chiabu.2019.104180] [Medline: 31521909]
25. Perron BE, Victor BG, Ryan JP, Piellusch EK, Sokol RL. A text-based approach to measuring opioid-related risk among families involved in the child welfare system. Child Abuse Negl. Sep 2022;131:105688. [doi: 10.1016/j.chiabu.2022.105688] [Medline: 35687937]
26. Saxena D, Moon ESY, Chaurasia A, Guan Y, Guha S. Rethinking "risk" in algorithmic systems through a computational narrative analysis of casenotes in child-welfare. Presented at: CHI '23: CHI Conference on Human Factors in Computing Systems; Apr 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544548.3581308]
27. Perron BE, Luan H, Victor BG, Hiltz-Perron O, Ryan J. Moving beyond ChatGPT: local large language models (LLMs) and the secure analysis of confidential unstructured text data in social work research. Res Soc Work Pract. 2024. [doi: 10.1177/10497315241280686]
28. Stoll D, Wehrli S, Lätsch D. Case reports unlocked: harnessing large language models to advance research on child maltreatment. Child Abuse Negl. Feb 2025;160:107202. [doi: 10.1016/j.chiabu.2024.107202] [Medline: 39689392]
29. Hanson B, Stall S, Cutcher-Gershenfeld J, et al. Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research. Nature New Biol. Nov 2023;623(7985):28-31. [doi: 10.1038/d41586-023-03316-8] [Medline: 37907636]
30. van Giffen B, Herhausen D, Fahse T. Overcoming the pitfalls and perils of algorithms: a classification of machine learning biases and mitigation methods. J Bus Res. May 2022;144:93-106. [doi: 10.1016/j.jbusres.2022.01.076]
31. Octoman O, Cox S, Arney F, Chong A, Tucker E. Narrative and fixed-field data: are we underestimating the risk of family and domestic violence? Child Abuse Rev. Jul 2023;32(4):e2811. [doi: 10.1002/car.2811]
32. Taherdoost H. What are different research approaches? Comprehensive review of qualitative, quantitative, and mixed method research, their applications, types, and limitations. J Manag Sci Eng Res. 2022;5(1):53-63. [doi: 10.30564/jmser.v5i1.4538]
33. Green BL, Ayoub C, Dym Bartlett J, et al. It's not as simple as it sounds: problems and solutions in accessing and using administrative child welfare data for evaluating the impact of early childhood interventions. Child Youth Serv Rev. Oct 2015;57(40-49):40-49. [doi: 10.1016/j.childyouth.2015.07.015] [Medline: 26744551]
34. Martin I. Hackers have uploaded thousands of malicious files to AI's biggest online repository. Forbes. 2024. URL: https://www.forbes.com/sites/iainmartin/2024/10/22/hackers-have-uploaded-thousands-of-malicious-models-to-ais-biggest-online-repository/ [Accessed 2025-07-15]
35. Lupariello F, Sussetto L, Di Trani S, Di Vella G. Artificial intelligence and child abuse and neglect: a systematic review. Children (Basel). Oct 6, 2023;10(10):1659. [doi: 10.3390/children10101659] [Medline: 37892322]
36. Cohen D. Data scientists targeted by malicious hugging face ML models with silent backdoor. JFrog. 2024. URL: https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/ [Accessed 2025-07-15]
37. Fluke JD, Tonmyr L, Gray J, et al. Child maltreatment data: a summary of progress, prospects and challenges. Child Abuse Negl. Sep 2021;119(Pt 1):104650. [doi: 10.1016/j.chiabu.2020.104650] [Medline: 32861435]
38. Zhou J, Zhang J, Wan R, et al. Integrating AI into clinical education: evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors. BMC Med Educ. Mar 19, 2025;25(1):406. [doi: 10.1186/s12909-025-06916-2] [Medline: 40108629]

39. Hatem R, Simmons B, Thornton JE. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus. Sep 2023;15(9):e44720. [doi: 10.7759/cureus.44720] [Medline: 37809168]
40. Evans H, Snead D. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. NPJ Digit Med. Apr 10, 2024;7(1):89. [doi: 10.1038/s41746-024-01093-w] [Medline: 38600151]
41. Hart LC, Viswanathan M, Nicholson WK, et al. Evidence from the USPSTF and new approaches to evaluate interventions to prevent child maltreatment. JAMA Netw Open. Jul 1, 2024;7(7):e2420591. [doi: 10.1001/jamanetworkopen.2024.20591] [Medline: 38976263]
42. Timmons AC, Duong JB, Simo Fiallo N, et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspect Psychol Sci. Sep 2023;18(5):1062-1096. [doi: 10.1177/17456916221134490] [Medline: 36490369]
43. Gillingham P. Predictive risk modelling to prevent child maltreatment and other adverse outcomes for service users: inside the "black box" of machine learning. Br J Soc Work. Jun 2016;46(4):1044-1058. [doi: 10.1093/bjsw/bcv031] [Medline: 27559213]
44. Schoech D, Jennings H, Schkade LL, Hooper-Russell C. Expert systems: artificial intelligence for professional decisions. Comput Hum Serv. 1985;1(1):81-115. [doi: 10.1300/J407v01n01_06]
45. Hussey JM, Marshall JM, English DJ, et al. Defining maltreatment according to substantiation: distinction without a difference? Child Abuse Negl. May 2005;29(5):479-492. [doi: 10.1016/j.chiabu.2003.12.005] [Medline: 15970321]
46. Jent JF, Eaton CK, Knickerbocker L, Lambert WF, Merrick MT, Dandes SK. Multidisciplinary child protection decision making about physical abuse: determining substantiation thresholds and biases. Child Youth Serv Rev. Sep 1, 2011;33(9):1673-1682. [doi: 10.1016/j.childyouth.2011.04.029] [Medline: 21804681]
47. Manion K, Renwick J. Equivocating over the care and protection continuum: an exploration of families not meeting the threshold for statutory intervention. Soc Policy J N Z. 2008;33(70):94. URL: https://www.thefreelibrary.com/Equivocating+over+the+care+and+protection+continuum%3A+an+exploration...-a0181897147 [Accessed 2025-07-15]
48. Gillingham P. Decision support systems, social justice and algorithmic accountability in social work: a new challenge. Practice (Birm). Aug 8, 2019;31(4):277-290. [doi: 10.1080/09503153.2019.1575954]
49. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. Minds & Machines. Mar 2020;30(1):99-120. [doi: 10.1007/s11023-020-09517-8]
50. Shaw J, Ali J, Atuire CA, et al. Research ethics and artificial intelligence for global health: perspectives from the global forum on bioethics in research. BMC Med Ethics. Apr 18, 2024;25(1):46. [doi: 10.1186/s12910-024-01044-w] [Medline: 38637857]
51. Qadhi SM, Alduais A, Chaaban Y, Khraisheh M. Generative AI, research ethics, and higher education research: insights from a scientometric analysis. Information. 2024;15(6):325. [doi: 10.3390/info15060325]
52. Beatty D, Masanthia K, Kaphol T, Sethi N. Revealing hidden bias in AI: lessons from large language models. arXiv. Preprint posted online on Oct 22, 2024. [doi: 10.48550/arXiv.2410.16927]
53. Holdsworth J. What is AI bias? IBM. 2023. URL: https://www.ibm.com/think/topics/ai-bias [Accessed 2025-07-18]
54. Dignum V, Penagos M, Pigmans K, Vosloo S. Policy guidance on AI for children. UNICEF Innocenti. URL: https://www.unicef.org/innocenti/reports/policy-guidance-ai-children [Accessed 2025-04-22]
55. Mahomed S, Aitken M, Atabey A, Wong J, Briggs M. AI, children's rights, & wellbeing: transnational frameworks. The Alan Turing Institute. 2023. URL: https://www.turing.ac.uk/sites/default/files/2023-12/ai-childrens_rights-_wellbeing-transnational_frameworks_report.pdf [Accessed 2025-04-22]
56. Karystianis G, Adily A, Schofield PW, et al. Surveillance of domestic violence using text mining outputs from Australian police records. Front Psychiatry. 2021;12:787792. [doi: 10.3389/fpsyt.2021.787792] [Medline: 35222105]
57. Bhattacharya A, Stumpf S, De Croon R, Verbert K. Explanatory debiasing: involving domain experts in the data generation process to mitigate representation bias in AI systems. Presented at: CHI 2025: CHI Conference on Human Factors in Computing Systems; Apr 26 to May 1, 2025; Yokohama, Japan. [doi: 10.1145/3706598.3713497]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by Sherif Badawy; peer-reviewed by ChengDa Zheng, Philip Gillingham; submitted 06.03.2025; final revised version received 04.06.2025; accepted 16.06.2025; published 24.07.2025

Please cite as:
Lukmanjaya W, Butler T, Cox S, Perez-Concha O, Bromfield L, Karystianis G

Leveraging AI to Investigate Child Maltreatment Text Narratives: Promising Benefits and Addressable Risks
JMIR Pediatr Parent 2025;8:e73579
URL: https://pediatrics.jmir.org/2025/1/e73579
doi: 10.2196/73579