Applications of Large Language Models in the Field of Suicide Prevention: A Scoping Review

Glenn Holmes, Biya Tang, Sunil Gupta, Svetha Venkatesh, Helen Christensen, Alexis Estelle Whitton

Submitted to: Journal of Medical Internet Research
on: June 11, 2024

Table of Contents

Applications of Large Language Models in the Field of Suicide Prevention: A Scoping Review

Glenn Holmes[1] PhD; Biya Tang[1] PhD; Sunil Gupta[2] PhD; Svetha Venkatesh[2] PhD; Helen Christensen[1] PhD; Alexis Estelle Whitton[1] PhD

[1]Black Dog Institute University of New South Wales, Sydney Randwick AU
[2]Applied Artificial Intelligence Institute Deakin University Melbourne AU

Corresponding Author:
Alexis Estelle Whitton PhD
Black Dog Institute
University of New South Wales, Sydney
Hospital Road
Randwick
AU

Abstract

Background: Prevention of suicide is a global health priority. Around 800,000 individuals die by suicide yearly, and for every death, there are another 20 estimated suicide attempts. Large language models (LLMs) hold the potential to enhance scalable, accessible, and affordable digital services for suicide prevention and self-harm interventions. However, their use also raises clinical and ethical questions that require careful consideration.

Objective: This scoping review aimed to identify emergent trends in applications of LLMs within the field of suicide and self-harm research. Additionally, it summarizes key clinical and ethical considerations relevant to this nascent area of research.

Methods: Searches were conducted in four databases. Eligible studies described the application of LLMs for suicide or self-harm prevention, detection, or management. English-language peer-reviewed articles and conference proceedings were included, with no date restrictions. This review adhered to PRISMA-ScR standards.

Results: Of the 533 studies identified, 36 met inclusion criteria, and an additional 7 more were identified through citation chaining, resulting in a total of 43 studies for review. A narrative synthesis approach was used to synthesize study characteristics, objectives, models, data sources, proposed clinical applications, and ethical considerations. Studies showed a bifurcation of publication fields with varying publication norms between computer science and mental health. While most studies (77%) focused on identifying suicide risk, newer applications leveraging generative functions (e.g., support, education, and training) are emerging. Social media was the most common source of LLM training data. BERT (Bidirectional Encoder Representation Transformer) was the predominant model used, although GPT (Generative Pre-trained Transformer) featured prominently in generative applications. Clinical applications of LLMs were reported in 60% of studies, often for suicide risk detection or as clinical assistance tools. Ethical considerations were reported in 33% of studies, with privacy, confidentiality, and consent strongly represented.

Conclusions: This evolving research area, bridging computer science and mental health, demands a multi-disciplinary approach. While open access models and datasets will likely shape this field, documenting their limitations and potential biases is crucial. High-quality training data is essential for refining these models and mitigating unwanted biases. Policies that address ethical concerns – particularly related to privacy and security when using social media data – are imperative. The emergence of generative AI signals a shift in approach, particularly in applications related to care, support, and education. Ongoing human oversight, whether through human-in-the-loop testing or expert external validation, is essential for responsible development and use.

(JMIR Preprints 11/06/2024:63126)
DOI: https://doi.org/10.2196/preprints.63126

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Review

Applications of Large Language Models in the Field of Suicide Prevention: A Scoping Review

Glenn Holmes, PhD[1], Biya Tang, PhD[1], Sunil Gupta, PhD[2], Svetha Venkatesh, PhD[2], Helen Christensen, PhD[1], Alexis E Whitton, PhD[1]

Author Affiliations
[1]Black Dog Institute, University of New South Wales, Sydney, NSW Australia
[2]Applied Artificial Intelligence Institute, Deakin University, Melbourne, VIC Australia

Corresponding Author
Alexis E Whitton, Ph.D.
Black Dog Institute, University of New South Wales
Hospital Road, Randwick, NSW, AUSTRALIA 2031
Ph: +61 2 9065 9046
E: a.whitton@unsw.edu.au

Word Count	
Abstract:	393
Main text: 8198	
Tables:	0
Figures: 3	

Abstract

Background

Prevention of suicide is a global health priority. Around 800,000 individuals die by suicide yearly, and for every death, there are another 20 estimated suicide attempts. Large language models (LLMs) hold the potential to enhance scalable, accessible, and affordable digital services for suicide prevention and self-harm interventions. However, their use also raises clinical and ethical questions that require careful consideration.

Objectives

This scoping review aimed to identify emergent trends in applications of LLMs within the field of suicide and self-harm research. Additionally, it summarizes key clinical and ethical considerations relevant to this nascent area of research.

Method

Searches were conducted in four databases. Eligible studies described the application of LLMs for suicide or self-harm prevention, detection, or management. English-language peer-reviewed articles and conference proceedings were included, with no date restrictions. This review adhered to PRISMA-ScR standards.

Results

Of the 533 studies identified, 36 met inclusion criteria, and an additional 7 more were identified through citation chaining, resulting in a total of 43 studies for review. A narrative synthesis approach was used to synthesize study characteristics, objectives, models, data sources, proposed clinical applications, and ethical considerations. Studies showed a bifurcation of publication fields with varying publication norms between computer science and mental health. While most studies (77%) focused on identifying suicide risk, newer applications leveraging generative functions (e.g., support, education, and training) are emerging. Social media was the most common source of LLM training data. BERT (Bidirectional Encoder Representation Transformer) was the predominant model used, although GPT (Generative Pre-trained Transformer) featured prominently in generative applications. Clinical applications of LLMs were reported in 60% of studies, often for suicide risk detection or as clinical assistance tools. Ethical considerations were reported in 33% of studies, with privacy, confidentiality, and consent strongly represented.

Conclusions

This evolving research area, bridging computer science and mental health, demands a multi-disciplinary approach. While open access models and datasets will likely shape this field,

documenting their limitations and potential biases is crucial. High-quality training data is essential for refining these models and mitigating unwanted biases. Policies that address ethical concerns – particularly related to privacy and security when using social media data – are imperative. The emergence of generative AI signals a shift in approach, particularly in applications related to care, support, and education. Ongoing human oversight, whether through human-in-the-loop testing or expert external validation, is essential for responsible development and use.

Keywords: suicide, suicide prevention, large language model, self-harm, artificial intelligence

Applications of Large Language Models in the Field of Suicide Prevention: A Scoping Review

Prevention of suicide is a global health priority [1]. Around 800,000 individuals die by suicide yearly, and for every suicide death there are 20 times as many people who have made a suicide attempt [1]. Although the wide-ranging impacts of suicide are considered largely preventable, factors such as limited service capacity, variable service quality, and barriers to service access significantly impact upon the progress being made towards reducing suicide rates [2]. Recent advances in transformer-based Artificial Intelligence (AI) – the technology that has accelerated development of powerful Large Language Models (LLMs) that human clinicians and consumers can converse with – have been suggested as a potential solution to enhancing the scalability, accessibility, and personalization of healthcare interventions [3]. In the context of suicide research and care, LLMs can generate novel insights into suicide risk by parsing language, classifying or scoring text, and comprehending and generating human-like language, with the potential to make treatments better by improving user engagement with digital interventions. Such improvements can also expand the capacity of crisis support services by providing real-time crisis clinician co-pilots, automated risk assessments, or expedited triage enabling optimized use of human resources; or to improve the quality of training programs for those who may intervene, by mentoring trainees through clinical scenarios and flexibly adopting a variety of crisis seeking personas for the purpose of clinical role-play scenarios. Hence, LLM technology might be a critical catalyst for advancement in the field. However, the relatively unexplored nature of the field means that we are only just beginning to understand how LLMs can be harnessed safely and effectively for suicide prevention. In parallel, there is a pressing need to explore how emerging trends in LLM-based research, such as the nature of training data, model types, contexts of application, and methodological norms from relevant research fields, may shape the trajectory and direction of suicide prevention research (either positively or negatively), now and into the future.

AI refers broadly to machine and computational processes used to execute tasks usually thought of as requiring human intelligence. Natural language processing, a language-focused subfield of AI, has advanced significantly with the advent of transformer-based LLMs [4]. LLMs are the current state-of-the-art technology in computational linguistics, comprising attention-based language encoders trained using massive text datasets, generally with billions of parameters [5]. In addition to language comprehension, LLMs have recently shown their utility in generative language applications, such as generative AI interfaces (chatbots) [6]. The availability of LLMs such as OpenAI's Generative pre-trained transformer (GPT) models, Google's Bard, and Meta's LLaMa has created unprecedented opportunities for language generation and analysis at scale [7], with applications already becoming widespread across fields such as business analytics, commerce, administration, and education [8].

Under the right conditions, LLMs can demonstrate contextual understanding and content generation that closely mimics human interaction. Accordingly, research into LLM applications in healthcare settings, where human interaction forms the basis of much of service delivery, is accelerating at a rapid pace. For example, LLMs have been investigated as a potential aid in pre-consultation, diagnosis, and management of disease [e.g., infectious diseases, cancer; 9, 10, 11], recommending specialist appointments via SMS-based self-assessment tools for remote populations [12] and by generating patient education materials [13].

Given that language is the primary basis upon which symptoms are reported and assessed in mental health, LLMs represent a significant technological advancement in mental health research. LLMs are currently under investigation for their utility in cognitive behavioral therapy facilitation [14]; for emotion identification during psychotherapy sessions [15]; and as a means to detect positive therapeutic behaviors during motivational interviewing [16]. LLMs have also shown potential benefits in helping individuals understand personal coping styles and in facilitating stress reappraisal [7]. Application of LLMs in the field of suicide prevention is particularly promising given that traditional methodologies, which have struggled to provide actionable insights into complex constructs such as suicide risk, are complemented by the analytic and generative capabilities of

LLMs [17]. Accordingly, LLMs hold promise for improving models of suicide risk (e.g., via digital phenotyping) [17]; for creating synthetic data [18] to increase sample sizes and statistical power, for research on low base-rate events such as suicide; for powering conversational agents that support early triage, crisis support, and help-seeking; and for clinician assistance tools such as automated patient assessment systems, and simulation of crisis scenarios to improve training outcomes. Despite these potentialities, the integration of LLMs into suicide prevention currently trails other fields. Critically, significant gaps must be addressed for the field to move forward in a manner that is safe and acceptable, including understanding the interpretability of LLM outputs, addressing biases within training data, and ensuring ethical deployment within prospective crisis settings.

To date, there have been no reviews examining the integration of LLMs in research on suicide prevention and self-harm. Prior reviews have focused predominantly on research undertaken in adjacent fields, such as applications of machine learning or chatbots to mental health more broadly [19-22]. Although several reviews have explored the use of machine learning [23-26] or AI-based strategies [27-30] in suicide prevention contexts, these have not focused on applications of LLMs specifically. Furthermore, although one commentary [31] and a review [24] focused on the integration of computational linguistics or natural language processing more broadly to suicide prevention, these reviews have not focused on LLMs specifically, nor have any explored applications of LLMs to self-harm.

Objective

The aim of this scoping review was to provide an understanding of current applications of LLMs in the field of suicide prevention, including applications to self-harm. Specifically, we aimed to identify emerging trends related to the types of training data and models used, the contexts in which they are applied, and their intended outcomes. Our secondary aim was to explore clinical applications of LLMs and identify ethical considerations that are crucial as the field progresses. A *scoping* review was selected to map the breadth of evidence across the varied fields from which LLM-based research originates, to identify key characteristics of studies, identify knowledge gaps, and inform future research, [32]. In doing so we aim to highlight emergent trends in the research and address critical clinical and ethical considerations to ensure safe, effective, and equitable research outcomes in this key field.

Method

This review is presented in line with the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) and was developed and carried out with reference to best practice methodological standards published by the Joanna Briggs Institute [33, 34]. In accordance with recommendations from the Joanna Briggs Institute [35], the protocol was pre-registered (Open Science Framework) and is publicly available [36].

Inclusion and Exclusion Criteria

To meet the eligibility criteria, studies needed to describe application of an LLM to the area of suicide or self-harm. Although various natural language processing models utilizing machine learning methods existed previously (e.g., Support Vector Machines, Bayesian Networks, Random Forest algorithms), for the purpose of this review we focus on contemporary LLMs, defined as computer engineered language models that use transformer-based neural network architecture [4]. Though LLM training parameters generally exceed 10B there is no formal consensus on parameter scale for LLMs [6]. In this review smaller (<10B parameter) models such as early incarnations of the Bidirectional Encoder Representation Transformer (BERT) and T5, were eligible for inclusion as these represent language models capable of contextual understanding. Studies describing use of AI, machine learning, or big data approaches in the absence of LLM methodology, were ineligible.

The domain of suicide prevention included (but was not restricted to) areas such as: ideation, planning, attempts, prediction, intervention, support, means restriction, gatekeeper training, and public awareness campaigns, as well as self-harm. Studies focusing on LLM applications to mental health or psychiatry that did not specifically mention suicide or self-harm were ineligible. Studies describing quantitative, qualitative, or mixed method designs were eligible. All studies involving human subjects were required to report ethics approval. Source documents were peer-reviewed journal articles or conference proceedings. Conference proceedings were included as they have become the dominant form of published research in computer science (encompassing LLMs) in recent years [37]. Conference proceedings were required to have a comprehensive methodology and results, sufficient for replication. Conference abstracts were excluded, as were other reviews and meta-analyses.

Studies employing the use of electronic health records (EHRs) were not included in this review. EHRs represent a specific type of corpus representative of individuals who are in contact with the health system. Given that a significant proportion of individuals experiencing suicidal ideation are not in contact with formal health services [38], models that draw primarily on EHR data may generate insights that do not generalize to the broader population of individuals who experience suicidality or self-harm. In addition, substantial work has focused on the use of EHRs for the prediction of suicide risk in recent years (25) with the potential for these studies to populate a standalone systematic review. This review sought to elicit an understanding of novel deployments of LLMs, particularly in the individual use context, and to propose future use possibilities or potentialities for support, treatment, or prevention. Studies not published in English were also excluded from review. There was no limitation on country of origin or publication date, however the advent of transformer architecture in 2017 naturally limited publications from prior years.

Search Strategy

Searches were conducted in four databases: PsycINFO, EMBASE, PubMed, and Institute of Electrical and Electronics Engineers (IEEE). An initial search was conducted in December 2023 (6/12/23). Following abstract and full text screening, the search was updated in February 2024 (16/2/2024) to ensure recency of the search prior to final text extraction. Additional studies identified via citation chaining were also included at this point.

The search strategy used index terms and free-text terms to cover two core themes: (1) LLMs and (2) suicide and/or self-harm. Individual database search strings are provided in Table S1 in Multimedia Appendix 1. Search results were imported into Covidence [39] review management software, which was used for abstract and full-text screening, and data extraction.

Study Selection and Screening Process

Reflecting prior systematic research [40], title and abstract screening were conducted by one author (GH) with a sample of studies reviewed by multiple authors during screening and extraction. In line with recommendations for ensuring reliability and bias mitigation [32], 20% of studies (n = 77) were randomly selected and reviewed by a second author (BT). Agreement was achieved for 71 of 77 studies (92%). Conflicts were discussed and resolved with input from a third author (AEW).

A similar process was adopted for full text screening, which was conducted by one author (GH), with 20% of full texts (n = 38) randomly selected for review a second author (BT). There was 92% (35/38) agreement between authors. Conflicts were discussed and resolved with a third author (AEW).

Data Extraction

The data extraction template was piloted with 10% of included studies (n = 4 studies) by two authors (GH & BT) in line with prior research [40, 41]. Minor refinements were made to the template to assist later synthesis. Data extraction from the 43 included articles was performed by one author (GH). The data extraction template is available in Table S2 in Multimedia Appendix 1.

Analysis

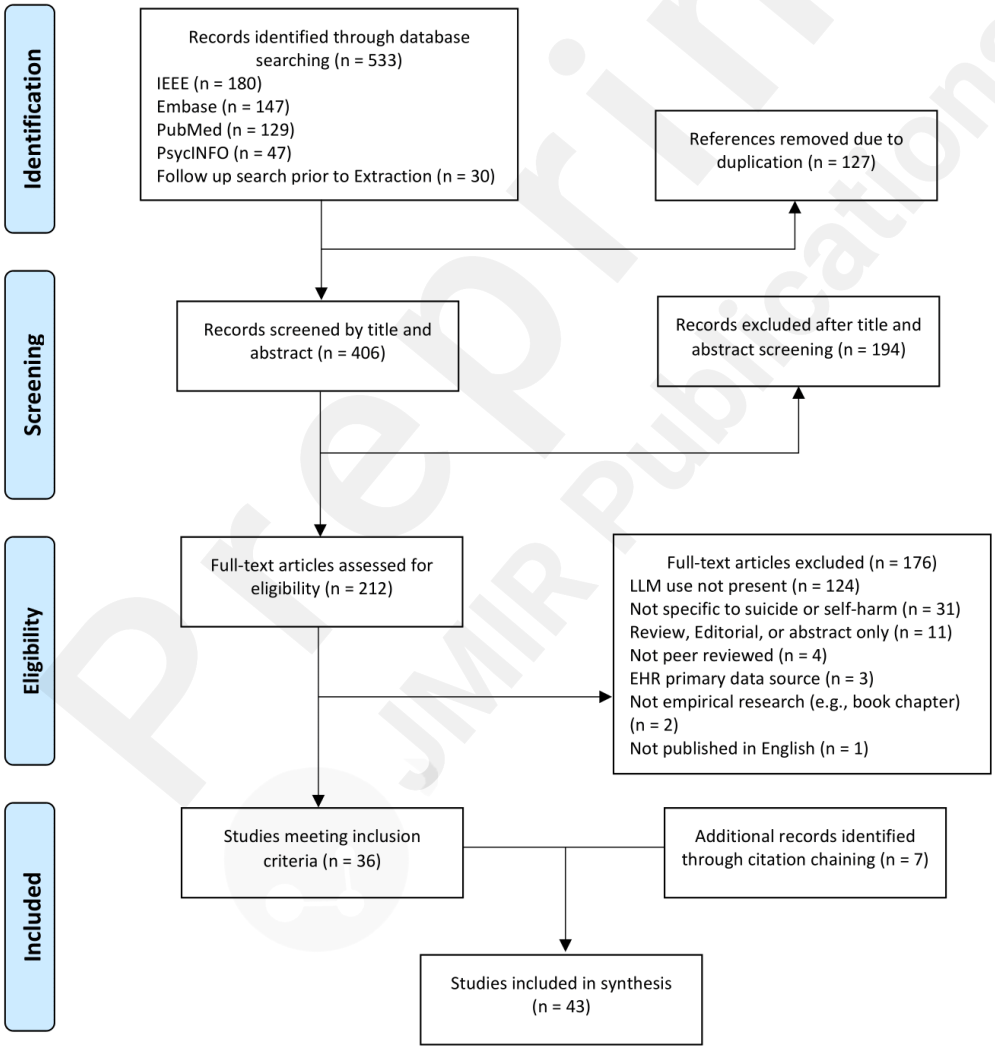
Studies were classified across deductively derived categories with categories based on previous research [41] with scope for adjustments during the review process if required (e.g., aggregation of specific data sources to report on social media more generally as a source of data). Descriptive statistics were computed and presented in text, table, or graphical format where appropriate. Narrative synthesis of included studies was used to answer the key research questions.

Results

Selection of Sources of Evidence

Search results are presented in the PRISMA flowchart (Figure 1). The search yielded 533 studies. Following de-duplication (127 duplicates), 406 were retained for title and abstract screening, and of these, 194 (48%) were excluded, leaving 212 (52%) for full-text review. After full-text screening, 176 (43%) studies were excluded (reasons shown in Figure 1). Thirty-six (9%) studies met inclusion criteria, and an additional 7 were identified through citation chaining, resulting in 43 studies included for synthesis.

Figure 1. PRISMA flowchart



Characteristics of Included Studies

A full list of included publications and associated results data is presented in Multimedia Appendix 2. Publications by year demonstrated a compounding trend, with initial publications emerging in 2019 following the development of the transformer architecture [4] in 2017, and then increasing sharply in

2023 following the release of ChatGPT to the public in November 2022.
Articles identified for this review came from 20 different sources that were distributed relatively equally across two distinct fields; Computer Engineering (n = 23; 53%) and Health (n = 18; 42%). The bifurcated nature of publication field-of-origin demonstrates the cross disciplinary nature of LLM research. About a third of studies were published as conference proceedings via IEEE (n = 16; 37%), reflecting publication norms in computer engineering [37]. The largest proportion of studies were led by authors from the Unites States of America (n = 10; 23%). Study funding was predominantly grant sourced (n = 23, 53%), 4 studies (9%) indicated no funding and 16 (37%) did not provide funding information.

Synthesis of Results

Base Large Language Model Utilized

A synthesis of identified trends with associated recommendations is presented in Figure 2. Most studies (n = 35; 81%) applied some derivation of Google's BERT. OpenAI's GPT family of LLMs were used in 9 (21%) studies. Other models included XLNet (n = 3), Google's Fine-tuned LAnguage Net (FLAN) (n = 2; 5%), Alpaca, Alexa, DeepMoji, and Contrastive Language-Image Pre-training (CLIP; n = 1 each; 2%). Some studies applied more than one model type. The widespread adoption of BERT can be attributed in part to its status as one of the earliest open-source models (accessible since its inception in 2018), allowing users to freely download and utilize the model for research purposes. FLAN [42] is also open source, though more recently released in 2022. CLIP, DeepMoji, Alexa, and GPT (3.5) are accessible via API but are not open-source models, limiting data transparency and reliability of availability required for most research applications.
In 33 (77%) studies, the primary purpose of LLMs was contextual understanding of language, while 9 (21%) studies extended upon this by using LLMs' capacity to generate natural language responses to prompts (all of which were published in 2023). One study also utilized an LLM for image interpretation [43]. Applications focused on contextual language understanding were predominantly employed for the purposes of identification, detection, or prediction of suicide risk. Generative applications also focused on prediction-based tasks [44-46], but extended to the evaluation of suicide risk [47-49], the identification of circumstances preceding suicide [50], information retrieval or question-answering systems [51], and creation of mental health nursing care plans [52]. Among the generative applications, seven out of nine utilized the text-based ChatGPT user interface, while the remaining two employed text-based interfaces for an educational BERT model [51] and a data secure FLAN model [50]. The prevalence of GPT use in generative applications stands in contrast to the dominance of the use of the BERT model across LLM-based research to date more generally. This hints at a potential shift away from studies primarily focused on BERT. Whether this trend reflects a broader shift in model preference or was influenced by factors such as availability, ease of use, or specific application requirements was not clear from the existing literature.

Data Sources

The majority (n = 31; 72%) of included studies utilized data consisting of user posts from social media platforms for LLM training, validation, testing, and deployment to answer a research question. Of these, Reddit was the most common platform from which data was derived (n = 18 studies), followed by Twitter/X (n = 13 studies). User posts from other platforms were also used [53-56]. Meta's Facebook was employed in only one study, wherein only images, not text data from users, were utilized [43]. Of the 31 studies employing social media data, user or poster consent for the use of their data was sought in only one study [43]. Of the 12 studies (28%) that did not use social media data, 3 used data from crisis counselling apps or services [57-59], with the remainder a heterogenous

collection of data from the National Violent Death Report System [50, 60], educational or academic documents [51, 61], or participant vignettes [49, 52]. One study utilized research participant data gathered during the course of the study [62]. One study did not use collected data but applied prompt inputs to ChatGPT to evaluate responses [48], and one study proposed an LLM moderated train safety device, designed to initiate braking and deploy an inflatable safety cell in front of the train following voice activation by the train driver [63].

Identified Objectives

Studies were grouped by four main objectives: Prediction; Identification/Classification; Support; and Education/Training. The majority of studies identified in this review applied LLM models to Identification/Classification tasks (n = 33; 77%), with prediction applications the next highest (n = 6, 14%). Studies focused on problems of Identification/Classification sought to identify content indicative of suicidal distress from text-based data, such as in Reddit posts [47, 64-71], Twitter/X posts [72-76], or in crisis or helpline conversations [57]. Additional uses included identifying precipitating events to suicide from death investigation narratives [60] and identifying self-harm from social media posts [77]. Prediction focused studies used LLMs utilizing Reddit [44, 78, 79], other social media [55], images posted on social media [43], crisis counselling data [58], or clinical data [78] predominantly to predict suicide risk.

The remaining four model applications assessed support [48], education [51], and ‘other’ (i.e., the generation of mental health care plans & means restriction) [52, 63]. Three of these four involved the use of generative AI. Regarding education, a question/answer interface was developed to generate specific information requested by individuals at risk and their families. The model underlying this interface drew on a corpus of over 300 suicide specific documents curated by clinicians [51]. Two studies used GPT, one with the aim of assessing the safety of publicly available conversational agents by prompting them with sequential PHQ9 items to examine chatbot responses to a patient simulation indicating escalating suicide risk [48]. The second, asked ChatGPT to generate mental health nurse care plans based on vignettes about a fictitious person self-harming [52]. Overall, study objectives focused on suicide and self-harm detection, while the development of generative LLM technologies has facilitated more recent LLM use in care, support, and education.

Clinical Applications

We examined trends in proposed clinical applications of LLMs to suicide or self-harm prevention. Studies were expected to have provided sufficient depth of discussion regarding ways in which the research could be translated into real-world settings. Brief general statements, or single sentence phrases about clinical applications were not considered sufficient for inclusion in addressing this research question.

Clinical applications were discussed in 26 studies (60%). Of the 17 publications that did not discuss clinical applications, the majority (n = 13; 76%) were from the field of computer science. Identified applications included improved detection of suicidality (n = 13), use as a clinical assistance tool (n = 10), improved accessibility of services (n = 3), improved services (n = 3), assisting the development of policy (n = 3), and use as a training tool (n = 2).

Specifically, studies discussing the ability for LLMs to improve detection of suicidality mentioned the creation of automated systems that could be applied to language data (e.g., social media posts) to detect suicidal ideation for the purposes of early intervention [57, 66, 80]. Studies focusing on the potential applications of LLMs as a clinical assistance tool discussed possible use cases where LLMs could aid clinicians in evaluating a client’s level of suicide risk [57, 66], supporting diagnosis and treatment [78, 81], providing a second opinion [49], or predicting a score on a mental health scale

[44]. Future oriented clinical assistance tool applications included AI-enabled avatars that could be remotely accessible, could deliver therapeutic services, could conduct simple examinations, provide advice, or recommend referrals for additional care [49]. Applications relating to service improvement included LLM integration to improve crisis counseling services [57, 59] and improving existing mental health chatbots [77]. Applications to training included enhancing training, clinical procedures, and best practices among mental health and medical professionals [59, 82]. Improved policy was noted as of clinical relevance [59, 82], as there may be a lack of policy safeguarding vulnerable people and the use of LLMs [52]. Cost effectiveness and increasing the quality of data annotation were also noted [47], as was use in public health surveillance, potentially allowing practitioners to track the prevalence of infrequent conditions [50].

Ethical Considerations

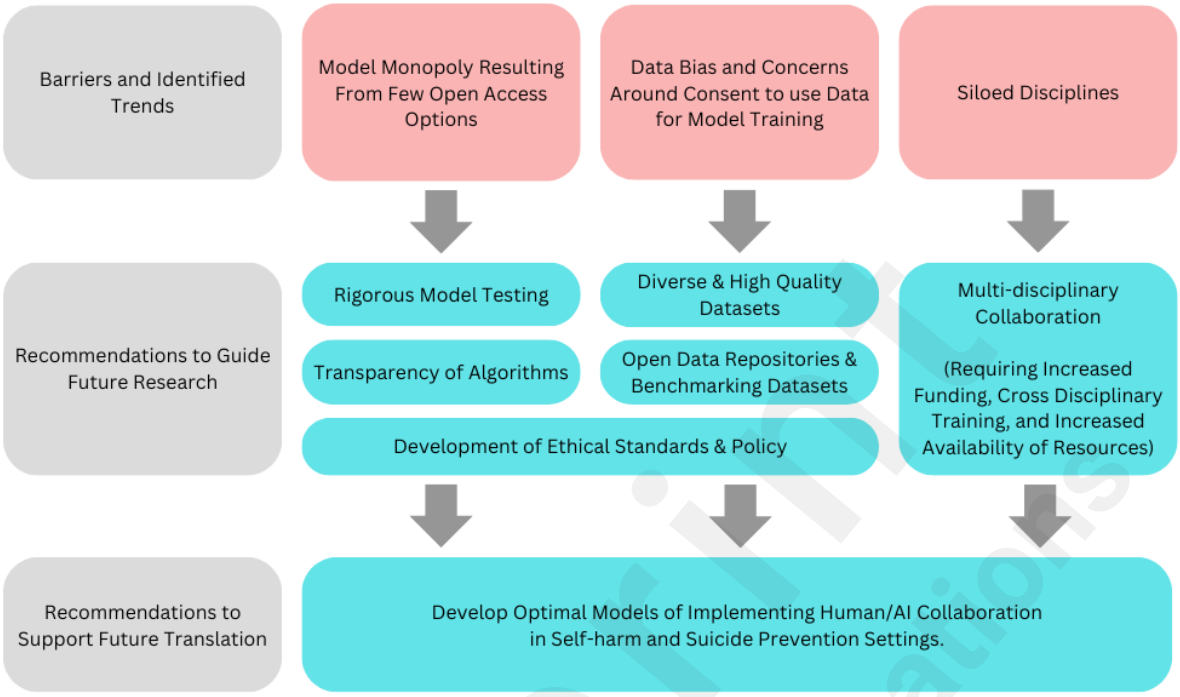
We also explored trends in the nature and type of ethical considerations reported. To be considered as having meaningfully reported on ethical issues, studies needed to include discussion of the ethical considerations and their possible implications. Cursory mention of potential ethical issues without discussion of their implications was not considered sufficient (e.g., mention of LLMs as being a 'black box' without any further discussion of the implications [43, 58] was not considered sufficient). Ethical considerations were discussed in 14 (33%) studies. Of the 29 (67%) studies not presenting ethical perspectives, the majority (n = 18; 62%) were published in computer engineering journals, where greater emphasis was given to the technical aspects of model development, training and validation. Privacy (n = 9) was the most common ethical consideration discussed, followed by bias (n = 5), hallucinations (n = 3), the 'black box' nature of LLMs (n = 3), possible threats to the client-clinician relationship (e.g., clinicians replaced with digital agents) (n = 2), and false positives/negatives in classification/identification (n = 2).

Specifically, studies that discussed ethical issues regarding privacy touched on concerns about confidentiality and consent to use publicly available data [44, 56, 80, 83], with researchers highlighting the need to value the privacy, security, and anonymity of the original posters and end users [49, 66]. Some researchers presented ethical concerns relating to publicly traded companies inferring or collecting sensitive health information about their users and acting on it, or sharing it, without explicit consent [47, 78]. Non-transparent data collection and inference processes currently used by social media platforms were highlighted as a growing area of concern [84]. Hallucinations (the tendency for LLMs to sometimes generate false responses), were identified as having impacts for the safety and reliability of LLM applications [47, 50]. Some studies raised concerns that models trained on vast amounts of data containing social biases or prejudices may perpetuate stigma towards suicide or self-harm [44, 47]. Related concerns were raised regarding the generalizability of LLMs to populations whose racial, ethnic, or cultural demographics differ from those present in training data [57, 83]. One study noted that these data biases or lack of diversity could lead to erroneous predictions that may exacerbate existing disparities in suicide prevention strategies [49]. Three studies noted lack of transparency, resulting from the opaque functioning of the underlying algorithms (termed 'black box') [49, 58], with one noting a lack of governance in this regard [52].

Two studies raised concerns about LLMs threatening the client-clinician relationship, potentially leading to human-centered clinical interactions being replaced by digital alternatives. Specifically, some studies discussed the issue of LLM agents lacking human attributes, such as empathy for suicidal distress [48]. Others noted that integration of LLMs into clinical care may lead to a sense of distancing of the clinician from the individual, potentially fostering feelings of invalidation or insignificance [52] and exacerbating suicidal thoughts or self-harm behaviors. False negatives (when suicidality goes undetected) and false positives (when suicidality is incorrectly flagged as being present) were noted as concerns [83, 84] as psychological harm can result (e.g., resulting in missed

opportunities to intervene with someone at risk, or in unnecessary mental health evaluations for someone who is not at risk) [84]. Safety was also noted as a concern as conversational AI may advance at a pace that outstrips associated safety measures [48]. Relatedly, issues of clinical responsibility were highlighted, particularly regarding the use of LLMs in aiding the generation of mental health care plans, as some authors believed this could leave mental health practitioners legally vulnerable [52]. More generally, studies emphasized that guidelines for safe development, auditing, and regulation are very much needed to address ethical risks in this area of research [44].

Figure 2. Trends and associated recommendations for ensuring safe and effective integration of LLMs into suicide prevention and self-harm research.



Discussion

The primary aim of this scoping review was to summarize and characterize emerging trends in the application of LLMs in the field of suicide prevention and self-harm research. This review maps the study characteristics and key methodological components of the 43 included studies, and further addresses the secondary aims of examining the clinical applications and ethical considerations proposed in the included studies.

Study Characteristics

The studies included in this analysis exhibited a notable divergence in disciplinary focus, with approximately equal representation from computer engineering (23) and health-related (18) fields. The bifurcation of domain expertise in this emerging field has important implications for the safety, effectiveness, and real-world impact of the research outputs, particularly considering the differences in technical training, publication norms, and approach to validation and evaluation, used across the different fields. For example, only recently has there been a call to integrate education in the technicalities and ethics of AI into mental health training programs, and many mental health researchers do not yet receive technical training in AI to be able to develop and apply LLMs for clinical research purposes. In parallel, while computer science researchers have significant technical training to enable development of innovative AI models and access to industry partnerships that can support the scalability of new AI-based tools, the real-world usefulness of these innovations hinges on a deep understanding of the ethical and clinical context in which they are to be applied, of the facilitators and barriers to their use, and of the needs and priorities of end users – areas of expertise where clinical training is often paramount. Recognizing and addressing the complementary strengths and gaps in these divergent disciplines will be crucial for ensuring LLM innovations in suicide prevention are grounded in technical excellence and clinical acumen.

Similarly, dissemination of research from computer science often out-paces the dissemination of research from the field of mental health, owing to the stronger focus on conference proceedings with expeditious publication timelines. The accelerated dissemination of LLM-based suicide prevention research originating from computer science fields may mean that practices focused on innovation – a crucial benchmark of research impact in computer science (e.g., via patents) [85] – come to dominate early research advancements in this area [86]. Although innovation is crucial for achieving technological advancements, this innovation often comes at expense of comprehensive clinical validation [87] and there has been growing emphasis on the need for frameworks for validating new AI tools in health research [88].

Recognizing and navigating these disciplinary disparities is essential. Encouraging multi-disciplinary collaboration among experts from diverse backgrounds is likely to be critical for bolstering the quality, safety, and impact of research on LLMs for suicide prevention. This could be done by fostering the development of agile methods for disseminating validation research (such as through cross-disciplinary repositories and benchmarking datasets), establishing sound ethical and safety frameworks that cut across key disciplines (such as living guidelines), and facilitating ongoing education and cross-disciplinary training for researchers across both fields. Crucially, prioritizing and supporting studies led by multidisciplinary teams that rigorously assess safety and effectiveness is imperative for ensuring research efforts lead to broader societal benefits.

Base Large Language Models

There was minimal variation in the base LLM model applied in the identified studies. Most studies (81%) used either the base BERT model or some variation of a trained or fine-tuned BERT model [89]. The relatively early release date (2018), open-source availability, compact size of this model compared to subsequent models like GPT, and its adaptability through fine-tuning for diverse applications, are all factors assisting BERT's widespread use in this research area. Whilst open-access is critical for facilitating replication and democratizing access to this modern technology, an important consideration is that open-access LLMs trained on generalized datasets might inadvertently perpetuate unwanted biases. For example, BERT has been shown to exhibit stereotypical biases in areas such as gender, profession, race, and religion, with this bias not related to size of pretraining corpora, but likely due to the nature of the training data [90]. In the same study GPT-2 was found to have less bias, hypothesized in part to be due to anti-stereotypical associations in the training data. LLMs are 'stochastic parrots' [91] in their generative responses, their outcomes depend on the quality of the input training data. Therefore, testing, application, and comparison of multiple models is important to detect and mitigate model biases permeating an entire discipline. Promoting use of a diverse range of models to assess performance variability [44, 47] and uncover potential biases [90] could enhance the strength of research outcomes in suicide prevention. Additionally, there is a pressing need to develop specialized open-access models (such as those being developed in medicine) [92, 93] for mental health and suicide prevention. These models should draw upon the expertise of mental health professionals, cross-disciplinary researchers, ethicists, and potentially individuals with lived experience. Establishing standards and policies that ensure transparent reporting of model training data and inherent potential biases is also imperative to enable critical evaluation and validation of model suitability [7], particularly for self-harm and suicide prevention contexts.

Data Sources

Most studies (31 out of 43) used data from social media platforms like Twitter/X and Reddit to train LLMs. Although these are rich data sources, a reliance on social media data as a form of training data may lead to limitations in the usefulness of LLMs. To protect their users, social media platforms often implement policies to detect, suppress, or remove content related to suicide or self-harm. Using social media data for LLM training may therefore result in an under-representation of critical data that would enable LLMs to effectively recognize and respond to expressions of suicidal distress. Additionally, algorithms embedded within social media platforms are designed to promote engaging content. As a result, these algorithms often promote user posting that is biased towards certain viewpoints and controversial or sensational topics, while simultaneously minimizing diversity of perspective. Accordingly, the algorithms in social media platforms may drive artificial patterns in the nature and frequency of certain types of user post data that do not reflect real-world interpersonal interactions.

It is also crucial to evaluate whether the profiles and behavior of individuals who post on social media platforms aligns well with the potential end-users of LLM-based suicide prevention interventions. For example, Reddit's user anonymity may constitute a perceived benefit to users, however, evidence indicates that users engage in differing behavior when posting anonymously compared to when their identity is known [94]. To mitigate this, it is imperative to curate training datasets that are diverse and balanced [95], and representative of the population the LLM is intended to serve [7]. Actively seeking out content, including positive interactions, supportive conversations, and safe discussions around self-harm and suicide prevention, may help enhance the effectiveness of

LLM-based suicide prevention interventions.

Identified Objectives

The studies included in this review predominantly deployed LLMs for the purposes of identification, detection, classification, and prediction of suicide or self-harm risk. These areas represent a narrow band of potential use, and despite research focused in this area there remains a lack of solutions that implement the models in clinical settings [80]. One reason for this is that computer scientists, who are at the forefront of technological innovation in this area, do not generally have in-depth knowledge of clinical processes and workflows used in the mental health field. This limits their ability to envision the broader applicability of LLMs in mental health settings, and results in a tendency to gravitate towards more familiar applications, such as classification and regression. Similarly, suicide prevention researchers often do not possess the technical expertise required to adapt LLMs for their specific needs. Consequently, multi-disciplinary collaborations are critical to the development of sound research programs with the clinical and technical expertise necessary for high caliber research. As this research provides further support and models continue to improve, models focused in clinical areas may be deployed in various contexts. However, it has been noted that this may be insufficient and may not reduce suicide attempts or deaths unless the treatment needs of people identified as being at risk are met [96]. Hence, it is imperative to explore innovative solutions in clinical care, consultation, and therapy.

This review highlighted a small number of recent applications that move beyond identification and prediction by leveraging the generative features of LLMs. These applications included assessing whether LLMs could provide access to suicide-related educational information [51], whether they could detect escalating suicide risk and the need for human intervention [48], and whether they could assist in generating mental health nurse care plans for individuals who self-harm [52]. These studies showed that at present, LLMs can perform well on specific elements of these roles, including providing immediate access to accurate information relevant to suicide prevention [51], using health theory frameworks to construct detailed care plans with tangible goals for clients [52], and in generating training aids for mental health providers [52]. However, these studies also highlighted areas where LLMs currently fall short. Specifically, they showed that LLMs did not effectively detect and respond to signs of escalating clinical risk by referring into care [48]. Moreover, concerns were noted regarding how the accountability of healthcare providers may be impacted by the use of AI generated mental health care plans [52]. These examples highlight the critical role of human expertise in the deployment of LLMs in suicide prevention contexts, and an important area for future research will be to determine optimal approaches to implementing human-AI collaboration in suicide prevention settings, including through the use of co-pilot or human-in-the-loop systems that are beginning to be explored in other areas [97, 98].

Clinical Applications

The secondary aim of this scoping review was to consider the clinical applications and ethical considerations of LLMs in suicide prevention practice. Of the studies describing clinical applications of LLMs ‘enhanced detection of suicidality’ and use as a ‘clinical assistance tool’ were the most commonly presented clinical applications. The ability to identify and predict suicide has remained at near-chance levels for many years [23] and improvement upon this with the help of novel technologies such as LLMs can provide a valuable contribution to the field. However, accurate risk detection alone is insufficient, and must be paired with effective and scalable interventions [23]. LLMs’ use as a clinical assistance tool can assist in supporting clinical professionals by aiding with a myriad of tasks [49, 57, 62, 66]. Conception and design of these applications should start with the

end point in mind, to best plan for the translation of research results into implementation in a clinical setting. Critically, researchers should engage clinicians, health professionals, lived experience representatives, and other relevant stakeholders early in research project development, to ensure the research can meet the needs of end users. Clinical applications described in the studies include assisting clinical decision-making by providing a second opinion [49], as a clinician co-pilot [57], as a mental health triage tool [66], and assisting in the delivery of therapy [62]. However, these were hypothesized future applications [49, 57, 66] or pilot studies [62] only. More investment is required in the development of implementable clinical assistance tools, particularly in supporting the integration of multidisciplinary teams to collaborate on areas such as clinical training and support as well as identifying novel use cases with the potential to positively impact suicide prevention practice. Importantly, LLMs should not set out to replace clinical judgement, but rather to supplement it, aiding clinicians and health professionals to make more informed decisions [49] by enhancing access to information (e.g., suggestion of differential diagnoses, alternate hypotheses, or collating scattered information into more coherent forms to facilitate human interpretation). LLM applications should not exceed this human-in-the-loop support based role until safety and effectiveness can be clearly demonstrated.

It is also important to consider the context where an LLM-based application will be implemented, acknowledging that its success may be limited by existing clinical workflows [99] and the overall capacity of the health system to integrate new tools and provide relevant training [100]. As with all new technological advances in care, LLM integration into healthcare practices requires strategic integration into existing, often outdated and practically constrained care models, if we are to realize their full benefit. Forward thinking and responsive policy is required, led by Governments that engage with stakeholders to map policy and standards to provide for the harmonious integration of LLM clinical support tools, with the accelerating use and potential of AI.

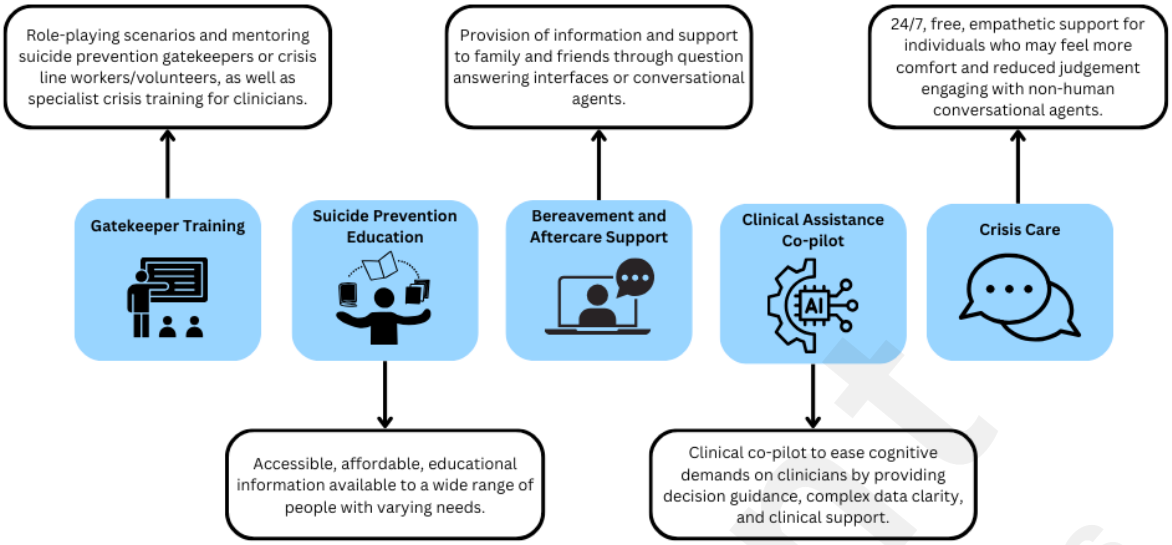
Research indicates that a majority of individuals do not engage with formal mental health services prior to suicide, potentially due to stigma or fear of judgement [38]. Whilst none of the identified studies deployed LLMs in support or intervention directly with distressed individuals, the ease of access that a language-based user interface offers may offer some individuals a means for overcoming stigma-related barriers to engaging with formal mental health services. Indeed, preliminary evidence for the acceptability of generative LLMs in support contexts has been demonstrated. For example, evidence shows that a large majority of individuals (78%) would be willing to use ChatGPT for the purpose of self-diagnosis or to aid in self-managing their symptoms [101]. Some individuals indicate a willingness to disclose to non-human agents citing less fear, reduced impression management, and increased ease expressing the severity of emotions [102].

This review found evidence that LLMs can deliver accurate information relevant to suicide prevention [51], and can support clinical needs in related contexts such as mental health evaluations, therapeutic consultations, and patient education [103]. LLMs therefore hold potential to offer individuals quick, empathic responses to mental health queries [104], whilst avoiding potential or perceived judgement and stigma, at little to no cost, 24 hours a day, 7 days a week. Yet, the efficacy of LLMs in providing precise, evidence-based support to individuals in suicidal distress has not been empirically validated. Furthermore, LLMs currently lack the ability to collect information vital for diagnosing and managing a patient's health condition [105]. For example, ChatGPT's performance has been found to deteriorate as the complexity of clinical cases increase [104], and research identified in this review demonstrates GPT-powered conversational agents and chatbots can be dangerously slow to escalate high-risk mental health situations for human clinical intervention [48]. While there may be a safe usage threshold where LLMs are beneficial for information and psychoeducational purposes, they are not advanced enough to be used as standalone therapeutic devices. Looking forward, models of LLM use in crisis support contexts are most likely to be those that are driven by clinical oversight, and that have sufficient guardrails in place to trigger referrals when clinically indicated.

The incorporation of LLMs into training and educational contexts presents a promising opportunity for research and development under clinical oversight. Educational resources, such as the question answering system in an included study that was aimed at providing suicide prevention information [51], provide an illustrative example. This mode of education is highly scalable and may provide a means for deploying training programs in a cost-effective manner. Such educational tools would be particularly useful in large scale, multi-faceted community wide interventions such as the European Alliance Against Depression [106] and Lifespan [107] models, both of which have community awareness and training as key components. Also identified in the review, Woodnutt and colleagues [52] suggest that LLMs like ChatGPT could support training, assisting less experienced mental health practitioners in creating care plans, brainstorming ideas, or pinpointing relevant aspects of patient presentation. However, it was noted that these generated plans, while appearing credible to laypersons, have been found to contain significant errors and ethical issues upon professional evaluation, highlighting the indispensable role of clinical oversight [52]. LLMs could also facilitate the creation of diverse case scenarios for healthcare providers to hone risk assessment and communication skills. By role-playing case scenarios, LLMs can offer real time, interactive training experiences for crisis call centre staff, counsellors, and suicide prevention gatekeepers, enhancing their confidence in discussing suicide and building clinical competencies. Additionally, the use of LLMs to converse in a roleplay scenario has the potential to improve initial training outcomes [108]. Critically, suicide prevention gatekeeper training outcomes are shown to diminish over time [109], however, engagement with a roleplaying LLM could result in improved retention of training outcomes and more at-risk individuals identified, approached, and referred for help.

These training and education use cases should be moderated by human oversight and iterated with human feedback for continuous and rapid improvement. Future research should aim to develop models trained in this domain that can demonstrate reliability and effectiveness in providing training or educational outputs. Development of the underlying model in this use case has the potential to translate to alternate clinical applications whilst providing researchers with a greater understanding of the guardrails and parameters of operation that are required for safe deployment across suicide prevention applications.

Figure 3. Potential future applications of LLMs in the field of self-harm and suicide prevention



Ethical considerations

Growing emphasis has been placed on the need for a clear and comprehensive ethical framework for integrating LLMs into mental health research and practice [110]. Hence, another aim of this review was to understand the ethical issues involved in the integration of LLMs into suicide prevention and self-harm research. Ethical considerations raised in the included studies focused on privacy and autonomy, bias, transparency and trust, and potential adverse impacts on the therapeutic relationship.

The integration of LLMs and social media data was noted as presenting a double-edged sword in terms of balancing privacy and confidentiality concerning the use of publicly available data with opportunities for suicide detection [44, 56, 80, 83]. On the one hand, the vast repository of publicly available language-based data on social media platforms provides the potential for real time identification of individuals at risk. While such methods can be lifesaving, they bypass the consent traditionally required in research and clinical practice. User or poster consent was sought in only 1 of the 31 identified studies that used social media data for training their models. Users of social media sites are often not afforded the opportunity to opt out of such surveillance or analysis, nor are they made aware of the potential adverse consequences associated with having their data used in this manner (e.g., the risk of breaches to data privacy, as has happened with some LLMs) [111]. This contravenes the fundamental human right to privacy, recently underscored by European legislation prohibiting certain data gathering practices related to health information by the algorithms of Facebook and other large social media companies [84, 112]. The use of social media and other data to train LLMs requires deep discussion to discern the regional, social, cultural, and temporal factors (among others) that influence ethical decisions around safe use of this data. An important ethical concern for the field of LLM-based suicide prevention research is how to balance the potential safety of users with their fundamental right to privacy. Real-world clinical application of LLMs focused on detecting individuals at heightened suicide risk also raises critical ethical issues regarding the implications of false positives, if acted on by authorities, and the potential threats this poses to personal autonomy.

Bias is another important ethical consideration that was noted in the identified studies. In a recent review of AI algorithms applied to mental health, Straw and colleagues [113] found significant biases exist with respect to religion, race, gender, nationality, sexuality, and age. These biases result from the expression of data (the manner in which the original data is presented) [114], the analysis of data (influenced by the contextual pre-learnings of the model), and the interpretation of results (human

annotation influenced by unconscious bias may produce model bias following training on that annotated data set) [113]. The presence of bias at any stage of model development risks creating tools that disadvantage certain groups of individuals [113]. Rigorous multi-model testing, comparison, and validation is required to isolate and mitigate inherent biases, and thereby ensure unbiased performance across diverse populations. Understanding the specific factors that contribute to the effectiveness or limitations of a model also allows researchers to make informed decisions about how to optimize future training processes. This may involve collecting more diverse and representative data, implementing better preprocessing techniques, or fine-tuning model parameters.

Ethical issues of transparency and trust were also discussed in several of the included studies. In contrast to traditional machine learning algorithms, LLMs are opaque black box architectures with complex internal structures, which makes it difficult to understand and explain their decisions [81]. This is particularly challenging when hallucinations occur, with little recourse for discovering the source of the error. LLMs do not offer their 'reasoning' unless prompted, and when prompted often fail to articulate how the provided information was retrieved, vetted, curated, or prepared. This lack of transparency is a fundamental flaw in our ability to understand how these models operate, challenging their trustworthiness in healthcare applications [81]. Efforts are being made to distil this unknown 'thinking' with the application of explainable AI methods. Two of the studies included in this review [68, 81] applied such methods in the form of SHAP [115], LIME [116], and Topic BERT [117]. Results indicate that these techniques can provide reasonable explainability for both short and long user-generated text and give insights about data quality issues in training datasets [81]. At present methods in this area are somewhat short of delivering the transparency needed to establish total trust, limiting the areas of deployment that restrict or protect against the potential for model hallucinations. Ongoing research into explainable AI techniques is critical in bridging this transparency gap to facilitate trust and enable real world implementation.

Two studies noted the client-clinician relationship as an ethical consideration, one in relation to LLMs agents lacking human attributes [48] and the other noting the distancing of the clinician from the person potentially fostering feelings of invalidation or insignificance [52]. At present, LLMs lack the authenticity and relational aspects required for modern mental health care [52]. Though they can simulate empathy, it is important to consider the specific care needs of individuals experiencing suicidal ideation. If deployed incorrectly, LLM technologies have the potential to incite harm [118], requiring careful implementation to maintain safe practices. When consideration is given to the development of LLMs for use in suicide prevention, sound policy needs to be written to safeguard care recipients, reflecting the complexity of the relationships they have with traditional care providers [52]. While acknowledging the important ongoing role of human-led therapies, there is a clear need for further research into developing AI systems that can effectively integrate professionals into the loop. This would allow LLMs to serve as an aid in clinical settings, complementing rather than supplanting human practitioners.

Limitations

Some limitations must be kept in mind when interpreting the findings of this scoping review. First, categorisation of machine learning technologies in the literature was not always clear, sometimes rendering it challenging to discern whether a study used an LLM or other machine learning model. It is possible some relevant articles may have been excluded from the review. Secondly, the cross-disciplinary nature of the field and resulting publications, reporting detail varied between disciplines. Health related publications provided more clinical and ethical related information, whereas computer engineering publications were more likely to provide in-depth detail about the LLM model and training process. This made synthesis and comparison of certain study attributes challenging. For example, data preparation and input methods were not sufficiently covered in studies published in

health-related journals to allow meaningful synthesis. Thirdly, all data were extracted by a single author. Although data extraction was piloted by 2 authors to ensure consistency of approach at the outset, some data that were extracted were qualitative, potentially resulting in researcher bias. Fourthly, though not a limitation, this review was conducted in an area of research that is currently experiencing significant growth and development. Therefore, though the review represents the current status of the literature, it only provides a time-stamped representation of the field.

Conclusions

LLMs represent a promising avenue for enhancing scalability, accessibility, affordability, and personalization of tools in the field of suicide prevention and self-harm, however collaboration between computer science and mental health experts is essential to leverage the strengths of both disciplines effectively. This review identified a strong bias towards the use of BERT with the potential for inherent biases indicating a pressing need for rigorous model comparison and testing, alongside the curation of diverse training datasets to mitigate model bias effectively. Increasing generative LLM applications indicate promise for transformative applications in care, support, training and education, however clinical accountability remains crucial to ensure the responsible use of LLMs. Identified ethical considerations underscore the need for clear governance via the establishment of policies and standards to guide the integration of LLMs into clinical use. LLMs have significant implications for self-harm and suicide prevention, underscoring the need to support continued research and development in this domain.

Funding: This research is funded by the Medical Research Future Fund (MRFF), Australia (Grant 1200195). HC and AEW received funding support from a NHMRC Investigator Grant (HC Grant 1155614; AEW Grant 2017521). The funder had no additional role in the production of this review.

Conflicts:
No conflicts of interest.

Abbreviations:
AI: Artificial intelligence
BERT: Bidirectional Encoder Representation Transformer
CA: Conversational agent
CLIP: Contrastive Language-Image Pre-training
FLAN: Fine-tuned LAnguage Net
GPT: Generative Pretrained Transformer
LLM: Large language model
ML: Machine learning
NLP: Natural language processing

Multimedia Appendix 1:
Individual database search strings and data extraction template.

Multimedia Appendix 2:
Table of included publications with extracted data.

References

1. World Health Organisation. Preventing suicide: A global imperative. 2014.
2. Martinez-Ales G, Hernandez-Calle D, Khauli N, Keyes KM. Why are suicide rates increasing in the united states? Towards a multilevel reimagination of suicide prevention. In: Baca-Garcia E, editor. Behavioral neurobiology of suicide and self harm. Cham: Springer International Publishing; 2020. ISBN: 978-3-030-57574-8
3. Melia R, Francis K, Hickey E, Bogue J, Duggan J, O'Sullivan M, et al. Mobile health technology interventions for suicide prevention: Systematic review. JMU; 2020;8(1):e12516. PMID: 31939744. doi: 10.2196/12516.
4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst; 2017;30. doi: 10.48550/arXiv.1706.03762.
5. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. Health Care Science; 2023;2(4):255-63. doi: 10.1002/hcs2.61.
6. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. ArXiv; 2023;abs/2303.18223.
7. Demszky D, Yang D, Yeager DS, Bryan CJ, Clapper M, Chandhok S, et al. Using large language models in psychology. Nat Rev Psychol; 2023;2:688-701. doi: 10.1038/s44159-023-00241-5.
8. Gado S, Kempen R, Lingelbach K, Bipp T. Artificial intelligence in psychology: How can we enable psychology students to accept and use artificial intelligence? Psychol Learn Teach; 2022;21(1):37-56. doi: 10.1177/14757257211037149.
9. He Y, Zhu Z, Zhang Y, Chen Q, Caverlee J. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. ArXiv; Preprint posted online October 8, 2020. doi: 10.48550/arXiv.2010.03746.
10. Li C, Zhang Y, Weng Y, Wang B, Li Z. Natural language processing applications for computer-aided diagnosis in oncology. Diagnostics; 2023;13(2):286. doi: 10.3390/diagnostics13020286.
11. Wang J, Zhang G, Wang W, Zhang K, Sheng Y. Cloud-based intelligent self-diagnosis and department recommendation service using chinese medical bert. J Cloud Comput; 2021;10:1-12. doi: 10.1186/s13677-020-00218-2.
12. Omoregbe NA, Ndaman IO, Misra S, Abayomi-Alli OO, Damaševičius R, Dogra A. Text messaging-based medical diagnosis using natural language processing and fuzzy logic. J Healthc Eng; 2020;2020:1-14. doi: 10.1155/2020/8839524.
13. Bala S, Keniston A, Burden M. Patient perception of plain-language medical notes generated using artificial intelligence software: Pilot mixed-methods study. JMIR Form Res; 2020;4(6):e16670. doi: 10.2196/16670.
14. Peretz G, Taylor CB, Ruzek JI, Jefroykin S, Sadeh-Sharvit S. Machine learning model to predict assignment of therapy homework in behavioral treatments: Algorithm development and validation. JMIR Form Res; 2023;7:e45156. PMID: 37184927. doi: 10.2196/45156.
15. Tanana MJ, Soma CS, Kuo PB, Bertagnolli NM, Dembe A, Pace BT, et al. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. Behav Res Methods; 2021;53(5):2069-82. doi: 10.3758/s13428-020-01531-z.
16. Shah RS, Holt F, Hayati SA, Agarwal A, Wang Y-C, Kraut RE, et al. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. Proc ACM Hum-Comput Interact; 2022;6(CSCW2):Article 527. doi: 10.1145/3555640.
17. Allen NB, Nelson BW, Brent D, Auerbach RP. Short-term prediction of suicidal thoughts and behaviors in adolescents: Can recent developments in technology and computational science provide a breakthrough? J Affect Disord; 2019;250:163-9. PMID: 30856493. doi: 10.1016/j.jad.2019.03.044.

18. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med; 2023;6(1):210. doi: 10.1038/s41746-023-00958-w.
19. Le Glaz A, Haralambous Y, Kim-Dufor D-H, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: Systematic review. J Med Internet Res; 2021;23(5):e15708. PMID: 33944788. doi: 10.2196/15708.
20. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. Can J Psychiatry; 2019;64(7):456-64. PMID: 30897957. doi: 10.1177/0706743719828977.
21. Xue J, Zhang B, Zhao Y, Zhang Q, Zheng C, Jiang J, et al. Evaluation of the current state of chatbots for digital health: Scoping review. J Med Internet Res; 2023;25:e47217. PMID: 38113097. doi: 10.2196/47217.
22. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. Next generation: Chatbots in clinical psychology and psychotherapy to foster mental health-a scoping review. Verhaltenstherapie; 2022;32:64-76. doi: 10.1159/000501812.
23. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: Applications and ethics. Behav Sci Law; 2019;37(3):214-22. doi: 10.1002/bsl.2392.
24. Arowosegbe A, Oyelade T. Application of natural language processing (nlp) in detecting and preventing suicide ideation: A systematic review. Int J Environ Res Public Health; 2023;20(2). PMID: 36674270. doi: 10.3390/ijerph20021514.
25. Burke TA, Ammerman BA, Jacobucci R. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. J Affect Disord; 2019;245:869-84. PMID: 30699872. doi: 10.1016/j.jad.2018.11.073.
26. Ji S, Pan S, Li X, Cambria E, Long G, Huang Z. Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems; 2020;8(1):214-26. doi: 10.1109/TCSS.2020.3021467.
27. D'Hotman D, Loh E. Ai enabled suicide prediction tools: A qualitative narrative review. BMJ HCI; 2020;27(3). PMID: 33037037. doi: 10.1136/bmjhci-2020-100175.
28. Khan NZ, Javed MA. Use of artificial intelligence-based strategies for assessing suicidal behavior and mental illness: A literature review. Cureus; 2022;14(7):e27225. PMID: 36035036. doi: 10.7759/cureus.27225.
29. Lejeune A, Le Glaz A, Perron PA, Sebti J, Baca-Garcia E, Walter M, et al. Artificial intelligence and suicide prevention: A systematic review. Eur Psychiatry; 2022;65(1):1-22. PMID: 35166203. doi: 10.1192/j.eurpsy.2022.8.
30. Bernert RA, Hilberg AM, Melia R, Kim JP, Shah NH, Abnousi F. Artificial intelligence and suicide prevention: A systematic review of machine learning investigations. Int J Environ Res Public Health; 2020;17(16):1-25. PMID: 2004922039. doi: 10.3390/ijerph17165929.
31. Ophir Y, Tikochinski R, Brunstein Klomek A, Reichart R. The hitchhiker’s guide to computational linguistics in suicide prevention. Clin Psychol Sci; 2022;10(2):212-35. doi: 10.1177/21677026211022013.
32. Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth; 2023;21(3):520-32. PMID: 02174543-202303000-00007. doi: 10.11124/jbies-22-00123.
33. Peters MDJ, Godfrey C, McInerney P, Munn Z, Tricco AC, Khalil H. Chapter 11: Scoping reviews (2020 version). 2020. In: JBI Manual for Evidence Synthesis [Internet]. Available from: https://doi.org/10.46658/JBIMES-20-12.
34. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. Prisma extension for scoping reviews (prisma-scr): Checklist and explanation. Ann Intern Med; 2018;169(7):467-73. PMID: 30178033. doi: 10.7326/m18-0850.

35. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth; 2020;18(10):2119-26. PMID: 02174543-202010000-00004. doi: 10.11124/jbies-20-00167.
36. Holmes G, Whitton, A., & Tang, B. Leveraging language based ai to improve suicide prevention: A scoping review. OSF; 2023. doi: 10.17605/OSF.IO/NCKQ7.
37. Mohammad SM. Examining citations of natural language processing literature. ArXiv; Preprint posted online May 2, 2020. doi: 10.48550/arXiv.2005.00912.
38. Tang S, Reily NM, Arena AF, Batterham PJ, Calear AL, Carter GL, et al. People who die by suicide without receiving mental health services: A systematic review. Frontiers in Public Health; 2022;9. doi: 10.3389/fpubh.2021.736948.
39. Covidence. Covidence review management. 2023. https://www.covidence.org/
40. Tang S, Werner-Seidler A, Torok M, Mackinnon AJ, Christensen H. The relationship between screen time and mental health in young people: A systematic review of longitudinal studies. Clin Psychol Rev; 2021;86:102021. PMID: 33798997. doi: 10.1016/j.cpr.2021.102021.
41. Harvey D, Lobban F, Rayson P, Warner A, Jones S. Natural language processing methods and bipolar disorder: Scoping review. JMIR Ment Health; 2022;9(4):e35928. PMID: 35451984. doi: 10.2196/35928.
42. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. JMLR; 2024;25(70):1-53.
43. Badian Y, Ophir Y, Tikochinski R, Calderon N, Klomek AB, Fruchter E, et al. Social media images can predict suicide risk using interpretable large language-vision models. J Clin Psychiatry; 2023;85(1):50516. PMID: 38019588. doi: 10.4088/jcp.23m14962.
44. Xu X, Yao B, Dong Y, Yu H, Hendler J, Dey AK, et al. Leveraging large language models for mental health prediction via online text data. ArXiv; 2023. doi: 10.1145/3643540.
45. Amin MM, Cambria E, Schuller BW. Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of chatgpt. IEEE Intell Syst; 2023;38(2):15-23. doi: 10.1109/MIS.2023.3254179.
46. Amin MM, Cambria E, Schuller BW. Can chatgpt's responses boost traditional natural language processing? IEEE Intell Syst; 2023;38(5):5-11. doi: 10.1109/MIS.2023.3305861.
47. Ghanadian H, Nejadgholi I, Osman HA. Chatgpt for suicide risk assessment on social media: Quantitative evaluation of model performance, potentials and limitations. ArXiv; 2023. doi: 10.48550/arXiv.2306.09390.
48. Heston TF. Safety of large language models in addressing depression. Cureus; 2023;15(12):e50729. doi: 10.7759/cureus.50729.
49. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of chatgpt-3.5 versus chatgpt-4: Vignette study. JMIR Ment Health; 2023;10:e51232. PMID: 37728984. doi: 10.2196/51232.
50. Zhou W, Prater LC, Goldstein EV, Mooney SJ. Identifying rare circumstances preceding female firearm suicides: Validating a large language model approach. JMIR Ment Health; 2023;10:e49359. PMID: 37847549. doi: 10.2196/49359.
51. Ascorbe P, Campos MS, Domínguez C, Heras J, Terroba-Reinares AR, editors. Towards a retrieval augmented generation system for information on suicide prevention. IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology; 2023; Malta: IEEE. doi: 10.1109/IEEECONF58974.2023.10404508.
52. Woodnutt S, Allen C, Snowden J, Flynn M, Hall S, Libberton P, et al. Could artificial intelligence write mental health nursing care plans? J Psychiatr Ment Health Nurs; 2023. PMID: 37538021. doi: 10.1111/jpm.12965.
53. Cao L, Zhang H, Feng L, Wei Z, Wang X, Li N, et al. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. Conference on Empirical Methods in Natural Language Processing and the 9th International Joint

Conference on Natural Language Processing (EMNLP-IJCNLP); Hong Kong, China, 2019. p. 1718-28. doi: 10.48550/arXiv.1910.12038.
54. Cao L, Zhang H, Wang X, Feng L. Learning users inner thoughts and emotion changes for social media based suicide risk detection. IEEE Trans Affect Comput; 2023;14(2):1280-96. doi: 10.1109/TAFFC.2021.3116026.
55. Yen S, Chu K, Tsai P, editors. Prediction model of social network suicide ideation by small sample. IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI); 2021 10-12 Aug. doi: 10.1109/IRI51335.2021.00060.
56. Wu EL, Wu CY, Lee MB, Chu KC, Huang MS. Development of internet suicide message identification and the monitoring-tracking-rescuing model in taiwan. J Affect Disord; 2023;320:37-41. PMID: 36162682. doi: 10.1016/j.jad.2022.09.090.
57. Broadbent M, Medina Grespan M, Axford K, Zhang X, Srikumar V, Kious B, et al. A machine learning approach to identifying suicide risk among text-based crisis counseling encounters. Front Psychiatry; 2023;14:1110527. PMID: 37032952. doi: 10.3389/fpsyt.2023.1110527.
58. Grimland M, Benatov J, Yeshayahu H, Izmaylov D, Segal A, Gal K, et al. Predicting suicide risk in real-time crisis hotline chats integrating machine learning with psychological factors: Exploring the black box. Suicide Life Threat Behav; 2024. PMID: 38345174. doi: 10.1111/sltb.13056.
59. Salmi S, Mérelle S, Gilissen R, van der Mei R, Bhulai S. Detecting changes in help seeker conversations on a suicide prevention helpline during the covid- 19 pandemic: In-depth analysis using encoder representations from transformers. BMC Public Health; 2022;22(1):530. PMID: 35300638. doi: 10.1186/s12889-022-12926-2.
60. Wang S, Dang Y, Sun Z, Ding Y, Pathak J, Tao C, et al. An nlp approach to identify sdoh-related circumstance and suicide crisis from death investigation narratives. J Am Med Inform Assoc; 2023;30(8):1408-17. PMID: 37040620. doi: 10.1093/jamia/ocad068.
61. Karapetian K, Jeon SM, Kwon JW, Suh YK. Supervised relation extraction between suicide-related entities and drugs: Development and usability study of an annotated pubmed corpus. J Med Internet Res; 2023;25:e41100. PMID: 36884281. doi: 10.2196/41100.
62. Mezzi R, Yahyaoui A, Krir MW, Boulila W, Koubaa A. Mental health intent recognition for arabic-speaking patients using the mini international neuropsychiatric interview (mini) and bert model. Sensors (Basel); 2022;22(3). PMID: 35161594. doi: 10.3390/s22030846.
63. Sheikh SA, Naidu H, editors. A novel robotics and mems artificial intelligence based train safety device. 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC); 2021 7-9 Oct. doi: 10.1109/ICOSEC51865.2021.9591761.
64. Bucur A-M, Cosma A, Dinu LP. Early risk detection of pathological gambling, self-harm and depression using bert. Conference and Labs of the Evaluation Forum (CLEF 2021); September 21-24; Bucharest, Romania. , 2021. doi: 10.48550/arXiv.2106.16175.
65. Dobbs MF, McGowan A, Selloni A, Bilgrami Z, Sarac C, Cotter M, et al. Linguistic correlates of suicidal ideation in youth at clinical high-risk for psychosis. Schizophr Res; 2023;259:20-7. PMID: 36933977. doi: 10.1016/j.schres.2023.03.014.
66. Garg M. Mental health analysis in social media posts: A survey. Arch Comput Methods Eng; 2023;30(3):1819-42. PMID: 36619138. doi: 10.1007/s11831-022-09863-z.
67. Haque F, Nur RU, Jahan SA, Mahmud Z, Shah FM, editors. A transformer based approach to detect suicidal ideation using pre-trained language models. 23rd International Conference on Computer and Information Technology (ICCIT) 2020; 2020 19-21 Dec. 2020. doi: 10.1109/ICCIT51783.2020.9392692.
68. Islam MR, Sakib MKH, Prome SA, Wang X, Ulhaq A, Sanin C, et al. Machine learning with explainability for suicide ideation detection from social media data. 10th International Conference on Behavioural and Social Computing (BESC); 30 Oct - 1 Nov: IEEE; 2023. p.

1-6. doi: 10.1109/BESC59560.2023.10386773.
69. Matero M, Idnani A, Son Y, Giorgi S, Vu H-H, Zamani M, et al. Suicide risk assessment with multi-level dual-context language and bert. The Sixth Workshop on Computational Linguistics and Clinical Psychology; Minneapolis, Minnesota 2019. doi: 10.18653/v1/W19-3005.
70. Naseem U, Khushi M, Kim J, Dunn AG. Hybrid text representation for explainable suicide risk identification on social media. IEEE Trans Comput Soc Syst; 2022:1-10. doi: 10.1109/TCSS.2022.3184984.
71. Sharma N, Karwasra P. Suicidal text detection on social media for suicide prevention using deep learning models. IEEE Region 10 Conference (TENCON); 31 Oct-3 Nov2023. p. 25-30. doi: 10.1109/TENCON58879.2023.10322499.
72. de Carvalho VcF, Giacon B, Nascimento C, Nogueira BM, editors. Machine learning for suicidal ideation identification on twitter for the portuguese language. Intelligent Systems; 2020 13 Oct; Cham: Springer International Publishing. doi: 10.1007/978-3-030-61377-8_37.
73. Ananthakrishnan G, Jayaraman AK, Trueman TE, Mitra S, A. A K, Murugappan A. Suicidal intention detection in tweets using bert-based transformers. International Conference on Computing, Communication, and Intelligent Systems (ICCCIS); 4-5 Nov; Greater Noida, India, 2022. p. 322-7. doi: 10.1109/ICCCIS56430.2022.10037677.
74. Metzler H, Baginski H, Niederkrotenthaler T, Garcia D. Detecting potentially harmful and protective suicide-related content on twitter: Machine learning approach. J Med Internet Res; 2022;24(8):e34705. PMID: 35976193. doi: 10.2196/34705.
75. Ravishankar TN, Kumar AK, Venkatesh J, Prabhu MR, Bhargavi VS, MuthamilSelvan.S. Empirical assessment and detection of suicide related posts in twitter using artificial intelligence enabled classification logic. International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI); 25-26 May; Chennai, India, 2023. p. 1-7. doi: 10.1109/ACCAI58221.2023.10201110.
76. Soudi RB, Zaghloul MS, Badawy OM, editors. Framework for suicide detection from arabic tweets using deep learning. 32nd International Conference on Computer Theory and Applications (ICCTA); 2022 17-19 Dec; Alexandria, Egypt. doi: 10.1109/ICCTA58027.2022.10206145.
77. Deshpande S, Warren J. Self-harm detection for mental health chatbots. In: Mantas Jea, editor. Medical Informatics in Europe (MIE2021); Virtual, 2021. p. 48-52. doi: 10.3233/SHTI210118.
78. Burkhardt HA, Ding X, Kerbrat A, Comtois KA, Cohen T. From benchmark to bedside: Transfer learning from social media to patient-provider text messages for suicide risk prediction. J Am Med Inform Assoc; 2023;30(6):1068-78. PMID: 37043748. doi: 10.1093/jamia/ocad062.
79. Lee DM, Moradi H, editors. Knowledge-infused dynamic embedding for predicting the severity of suicidal ideation in social media. 2022 International Conference on Computational Science and Computational Intelligence (CSCI); 2022 14-16 Dec. 2022. doi: 10.1109/CSCI58124.2022.00139.
80. Diniz EJ, Fontenele JE, de Oliveira AC, Bastos VH, Teixeira S, Rabêlo RL, et al. Boamente: A natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. Healthcare; 2022;10(4):698. doi: 10.3390/healthcare10040698.
81. Malhotra A, Jindal R. Xai transformer based approach for interpreting depressed and suicidal user behavior on online social networks. Cogn Syst Res; 2024;84:101186. doi: 10.1016/j.cogsys.2023.101186.
82. Wang S, Ning H, Huang X, Xiao Y, Zhang M, Yang EF, et al. Public surveillance of social media for suicide using advanced deep learning models in japan: Time series study from 2012 to 2022. JMIR; 2023;25. PMID: 2025131472. doi: 10.2196/47225.

83. Schoene AM, Bojanić L, Nghiem MQ, Hunt IM, Ananiadou S. Classifying suicide-related content and emotions on twitter using graph convolutional neural networks. IEEE Trans Affect Comput; 2023;14(3):1791-802. doi: 10.1109/TAFFC.2022.3221683.
84. Sawhney R, Joshi H, Nobles A, Shah RR. Towards emotion-and time-aware classification of tweets to assist human moderation for suicide prevention. Proceedings of the International AAAI Conference on Web and Social Media; 2021;15:609-20. PMID: 35173997. doi: 10.1609/icwsm.v15i1.18088.
85. Koya K, Chowdhury G. Measuring impact of academic research in computer and information science on society. 2nd Asia Pacific information technology conference, 2020. p. 78-85. doi: 10.1145/3379310.3379312.
86. Chubb J, Cowling P, Reed D. Speeding up to keep up: Exploring the use of ai in the research process. AI Soc; 2022;37(4):1439-57. PMID: 34667374. doi: 10.1007/s00146-021-01259-0.
87. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. npj Ment Health Res; 2024;3(1):12. PMID: 38609507. doi: 10.1038/s44184-024-00056-z.
88. Tsopra R, Fernandez X, Luchinat C, Alberghina L, Lehrach H, Vanoni M, et al. A framework for validating ai in precision medicine: Considerations from the european itfoc consortium. BMC Medical Informatics and Decision Making; 2021;21(1):274. doi: 10.1186/s12911-021-01634-3.
89. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv; Preprint posted online October 11, 2018. doi: 10.48550/arXiv.1810.04805.
90. Nadeem M, Bethke A, Reddy S. Stereoset: Measuring stereotypical bias in pretrained language models. ArXiv; Preprint posted online April 20, 2020. doi: 10.48550/arXiv.2004.09456.
91. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big?? Proceedings of the 2021 ACM conference on fairness, accountability, and transparency; March 1, 2021. p. 610-23. doi: 10.1145/3442188.3445922.
92. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. Pmc-llama: Toward building open-source language models for medicine. J Am Med Inform Assoc; 2024. doi: 10.1093/jamia/ocae045.
93. Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. ArXiv; Posted online May 19, 2023. doi: 10.48550/arXiv.2305.12031.
94. De Choudhury M, De S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. Proceedings of the international AAAI conference on web and social media; Cyprus, 2014. p. 71-80. doi: 10.1609/icwsm.v8i1.14526.
95. Chung NC, Dyer G, Brocki L. Challenges of large language models for mental health counseling. ArXiv; Preprint posted online November 23, 2023. doi: 10.48550/arXiv.2311.13857.
96. Kirtley OJ, van Mens K, Hoogendoorn M, Kapur N, de Beurs D. Translating promise into practice: A review of machine learning in suicide research and prevention. Lancet Psychiatry; 2022;9(3):243-52. PMID: 35183281. doi: 10.1016/s2215-0366(21)00254-6.
97. Ahmad MA, Yaramis I, Roy TD. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. ArXiv; Preprint posted online September 26, 2023. doi: 10.48550/arXiv.2311.01463.
98. Stade E, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. npj Ment Health Res; 2023;3(12). PMID: 38609507. doi: 10.1038/s44184-024-00056-z.

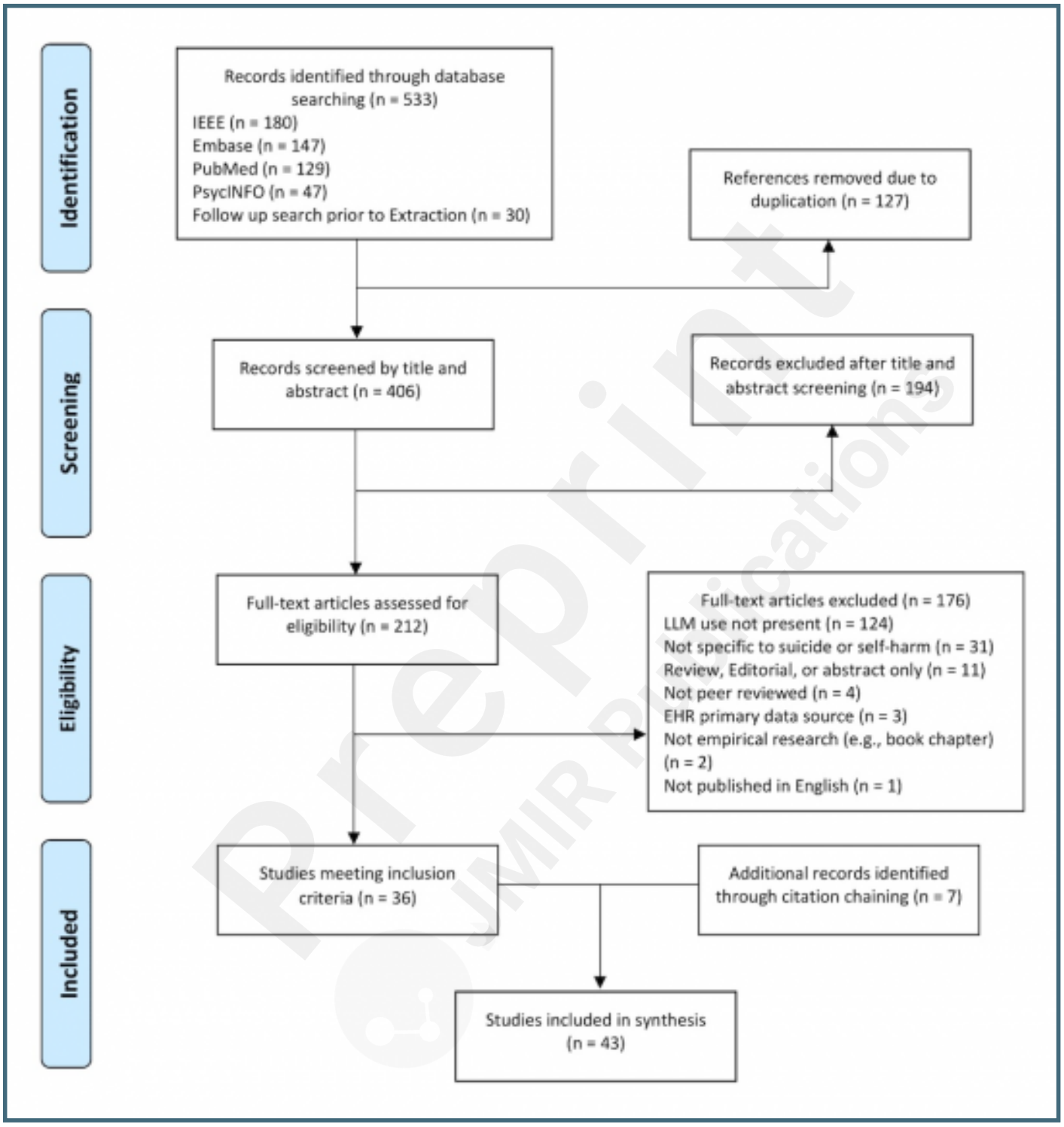
99. Laka M, Carter D, Milazzo A, Merlin T. Challenges and opportunities in implementing clinical decision support systems (cdss) at scale: Interviews with australian policymakers. Health Policy Technol; 2022;11(3):100652. doi: 10.1016/j.hlpt.2022.100652.
100. Jung K, Kashyap S, Avati A, Harman S, Shaw H, Li R, et al. A framework for making predictive models useful in practice. J Am Med Inform Assoc; 2021;28(6):1149-58. PMID: 33355350. doi: 10.1093/jamia/ocaa318.
101. Shahsavar Y, Choudhury A. User intentions to use chatgpt for self-diagnosis and health-related purposes: Cross-sectional survey study. JMIR Hum Factors; 2023;10(1):e47564. PMID: 37195756. doi: 10.2196/47564.
102. Lucas GM, Gratch J, King A, Morency L-P. It's only a computer: Virtual humans increase willingness to disclose. Comput Hum Behav; 2014;37:94-100. doi: 10.1016/j.chb.2014.04.043.
103. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of chatgpt in healthcare: An analysis of multiple clinical and research scenarios. J Med Syst; 2023;47(1):33. doi: 10.1007/s10916-023-01925-4.
104. Dergaa I, Fekih-Romdhane F, Hallit S, Loch AA, Glenn JM, Fessi MS, et al. Chatgpt is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry; 2024;14. PMID: 38239905. doi: 10.3389/fpsyt.2023.1277756.
105. Acuna Caicedo RW, Gomez Soriano JM, Melgar Sasieta HA. Bootstrapping semi-supervised annotation method for potential suicidal messages. Internet Interventions; 2022;28. PMID: 35281704. doi: 10.1016/j.invent.2022.100519.
106. Hegerl U, Althaus D, Schmidtke A, Niklewski G. The alliance against depression: 2-year evaluation of a community-based intervention to reduce suicidality. Psychol Med; 2006;36(9):1225-33. PMID: 16707028. doi: 10.1017/s003329170600780x.
107. Shand F, Torok M, Cockayne N, Batterham PJ, Calear AL, Mackinnon A, et al. Protocol for a stepped-wedge, cluster randomized controlled trial of the lifespan suicide prevention trial in four communities in new south wales, australia. Trials; 2020;21(1):1-10. PMID: 32293516. doi: 10.1186/s13063-020-04262-w.
108. Cross WF, Seaburn D, Gibbs D, Schmeelk-Cone K, White AM, Caine ED. Does practice make perfect? A randomized control trial of behavioral rehearsal on suicide prevention gatekeeper skills. J Prim Prev; 2011;32(3):195. doi: 10.1007/s10935-011-0250-z.
109. Holmes G, Clacy A, Hermens DF, Lagopoulos J. The long-term efficacy of suicide prevention gatekeeper training: A systematic review. Arch Suicide Res; 2021;2(25):177-207. PMID: 31809659. doi: 10.1080/13811118.2019.1690608.
110. McKernan LC, Clayton EW, Walsh CG. Protecting life while preserving liberty: Ethical recommendations for suicide prevention with artificial intelligence. Front Psychiatry; 2018;9:650. PMID: 30559686. doi: 10.3389/fpsyt.2018.00650.
111. Jang H. A south korean chatbot shows just how sloppy tech companies can be with user data2021 23/04/24. Available from: https://slate.com/technology/2021/04/scatterlab-lee-luda-chatbot-kakaotalk-ai-privacy.html.
112. Digital services act, Pub. L. No. 2022/2065 Stat. 32022R2065 (19 October 2022, 2022).
113. Straw I, Callison-Burch C. Artificial intelligence in mental health and the biases of language based models. PloS one; 2020;15(12):e0240376. PMID: 33332380. doi: 10.1371/journal.pone.0240376.
114. Wongkoblap A, Vadillo MA, Curcin V. Researching mental health disorders in the era of social media: Systematic review. J Med Internet Res; 2017;19(6):e228. PMID: 28663166. doi: 10.2196/jmir.7215.
115. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst; 2017;30.
116. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of

any classifier. 22nd ACM SIGKDD international conference on knowledge discovery and data mining; August 13, 2016. p. 1135-44. doi: 10.1145/2939672.2939778.
117. Grootendorst M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. ArXiv; Preprint posted online March 11, 2022. doi: 10.48550/arXiv.2203.05794.
118. Arora A, Arora A. The promise of large language models in health care. Lancet; 2023;401(10377):641. doi: 10.1016/S0140-6736(23)00216-7.

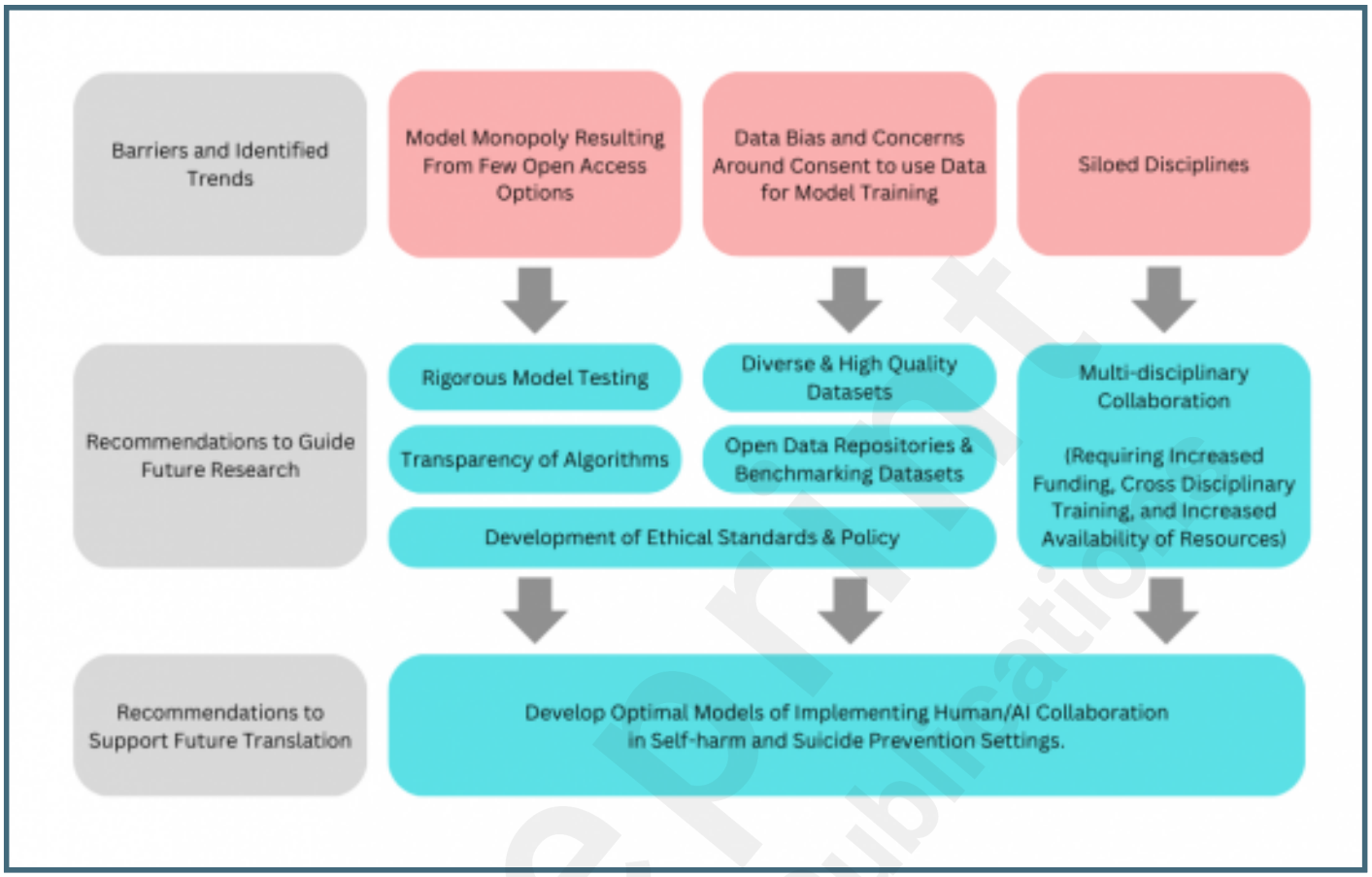
Supplementary Files

Figures

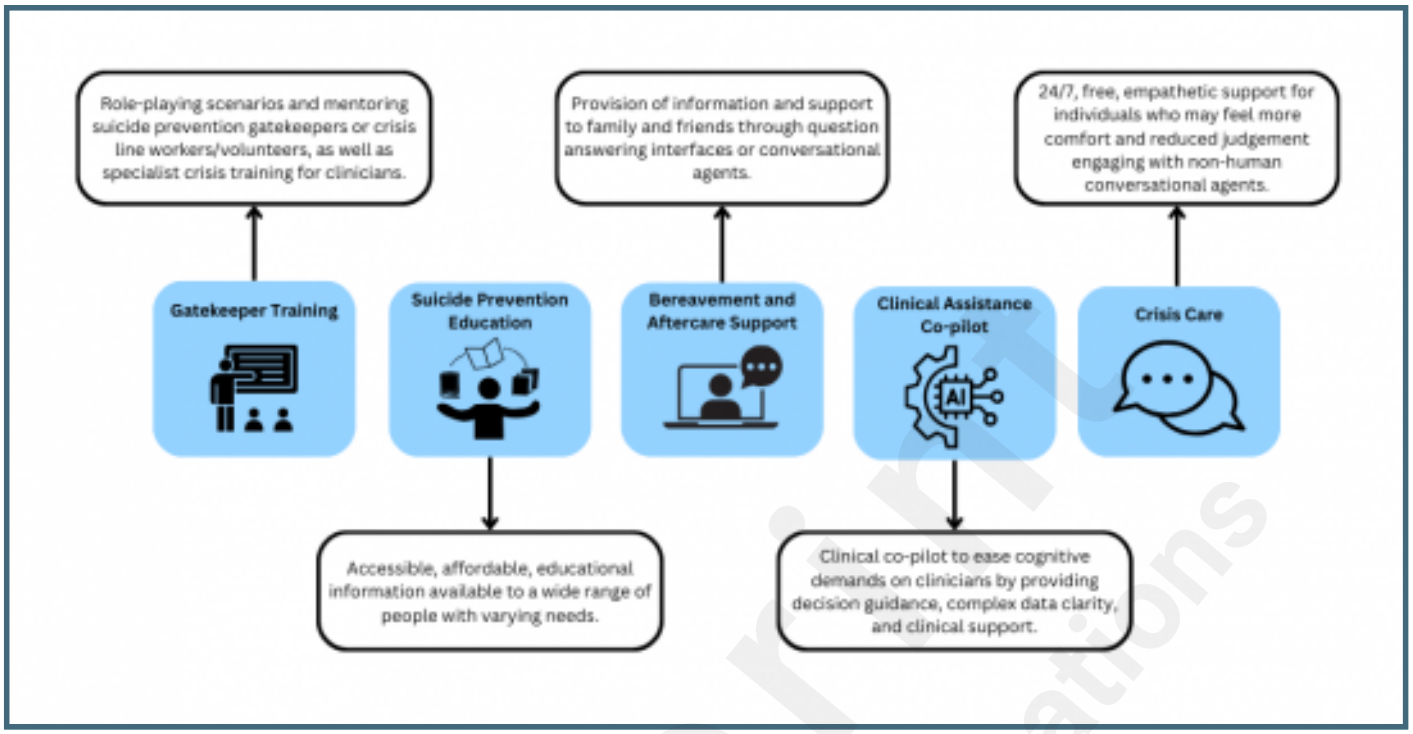
PRISMA flowchart.



Trends and associated recommendations for ensuring safe and effective integration of LLMs into suicide prevention and self-harm research.



Potential future applications of LLMs in the field of self-harm and suicide prevention.



Multimedia Appendixes

Search strings and data extraction template.
URL: http://asset.jmir.pub/assets/f1765587849d1141fa813a3b093e2824.docx

Extracted data table.
URL: http://asset.jmir.pub/assets/1aea90fd69fc019c7c9d21ff4e843016.xlsx

CONSORT (or other) checklists

PRISMA ScR checklist.
URL: http://asset.jmir.pub/assets/cc7f845fff25ef0f3aea2e071f377816.pdf